An Exploration of EEG-based, Non-stationary Emotion Classification for Affective Computing

by

Nicole Bendrich

Bachelors of Science in Engineering, University of New Brunswick, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Masters of Science in Engineering

In the Graduate Academic Unit of Electrical and Computer Engineering

Supervisor(s):	Erik J. Scheme, Ph.D., Electrical and Computer Engineering
Examining Board:	Kevin Englehart, Ph.D., Electrical and Computer Engineering, Chair
	Yevgen Biletskiy, Ph.D., Electrical and Computer Engineering
	Michael Fleming, Ph.D., Faculty of Computer Science

This thesis is accepted

by the

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

July, 2020

Abstract

The monitoring of emotional state is important in the prevention and management of mental health problems and is increasingly being used to support affective computing. Researchers are exploring various modalities from which emotion can be inferred, such as through facial images or via electroencephalography (EEG) signals. Current research commonly investigates the performance of machine-learning-based emotion recognition systems by exposing users to short films that are assumed to elicit a single known emotional response. Assuming static emotions, even for these brief periods, however, does not consider that emotions evolve. Moreover, in order to demonstrate better results, many existing models are not tested in ways that reflect realistic real-world implementations. In this thesis, the dynamic evolution of emotions induced using longer and variable stimuli is explored using EEG signals from the publicly available dataset, AMIGOS. A variety of feature engineering and selection techniques are applied and evaluated across four different cross-validation frameworks. The role of imperfect labelling of ground truth emotions and both data and gender-imbalances in the dataset are also investigated. Improved feature design and selection lead to up to 13% absolute improvement relative to comparable previously reported studies using this dataset. Alternative training configurations and a selective confidence-based classification scheme are proposed, leading to further possible improvements.

Acknowledgements

To my supervisor, Dr. Erik Scheme, thank you for providing me with opportunities to learn, topics to explore, and room to grow as a person and as a professional. For this, I will be forever grateful. Thank you to all the faculty, staff, and students at IBME. My time at IBME has been a positive experience because of all of you.

Thank you, Dr. Pradeep Kumar, for always taking the time to provide me with feedback and guidance as well as for your ceaseless positivity. To Dr. Dawn MacIsaac, thank you for introducing me to Python, and thank you, James Cameron, for encouraging me to use Python for my Master's work.

To my friends, Kathryn Cameron, Jena Nawfel, and Nitin Seth: thank you so much for the laughs and the adventures. My Master's has truly been a joy because of you all. To my family, thank you. Thank you for the ever-present support, no matter what trajectory I choose. I know I can do anything with your support. And finally, thank you to my other half, Ryan MacDonald. There are not enough words to describe how grateful I am, or how lucky I feel, to have your support.

Last but not least, I would like to thank the University of New Brunswick (UNB) and the Natural Sciences and Engineering Research Council (NSERC) for the continued support through my Master's degree.

Table of Contents

List of Tables

List of Figures

Nomenclature and Abbreviations

AF or AF3, AF4	Electrodes covering the anterior frontal lobe
Affect	Another word for emotion, thus it encompasses both valence and arousal
Arousal	A measure of stimulation
Asymm	Asymmetry
BVP	Blood Volume Pulse
CLF	Classifier
CNN	Convolutional Neural Network
CV	Cross-validation
DCNN	Deep Convolutional Neural Network
DFA	Detrended Fluctuation Analysis
ECG	Electrocardiography
EEG	Electroencephalography
ELM	Extreme Learning Machine
EMG	Electromyography
EOG	Electrooculography
F1-score	Classification Metric (Harmonic Mean of Recall and Precision)
F or F3, F4, F7, F8	Electrodes covering the frontal lobe
FACS	Face Action Coding System
FC or FC5, FC6	Electrodes covering the fronto-central lobe
FD	Fractal Dimension
FI	Fisher Information
FLD	Fisher's Linear Discriminant
fNIRS	Functional Near-Infrared Spectroscopy
FS	Feature Selection
GSR	Galvanic Skin Response
HA	High Arousal
HCI	Human-Computer-Interaction
HFD	Higuchi Fractal Dimension
HC	Hjorth Complexity
HM	Hjorth Mobility
HMM	Hidden Markov Model

HV	High Valence
Hz	Hertz
kNN	k-Nearest Neighbour
LA	Low Arousal
LDA	Linear Discriminant Analysis
LOMO	Leave-One-Movie-Out Cross-validation
LOO	Leave-One-Out Cross-validation
LOPO	Leave-One-Person-Out Cross-validation
LOPMO	Leave-One-Person-And-Movie-Out Cross-validation
LSTM	Long Short-term Memory
LV	Low Valence
MEG	Magnetoencephalography
mRMR	Minimum Redundancy Maximum Relevance
NB	Gaussian Naïve Bayes
NIR	Near-Infrared
O or O1, O2	Electrodes covering the occipital lobe
P or P7, P8	Electrodes covering the parietal lobe
PFD	Petrosian Fractal Dimension
PSD	Power Spectral Density
RBF	Radial Basis Function
RESP	Respiration
RF	Random Forest
RNN	Recurrent Neural Network
SBS	Sequential Backward Selection
SFS	Sequential Forward Selection
SVD	Singular Value Decomposition
SVM	Support Vector Machine
T or T7, T8	Electrodes covering the temporal lobe
TEMP	Skin Temperature
Valence	A measure of pleasantness
XGB	Extreme Gradient Boost

Chapter 1

Introduction

1.1 Motivation

According to the Mental Health Commission of Canada, 50% of Canadians have or will have had a mental health illness by the age of 40 [4], with mood and anxiety disorders among the most common types of mental health disorders [5]. Mood disorders can depress or elevate a person's mood whereas anxiety disorders are characterized by intense feelings of fear and distress [6]. Both disorders interfere with day-to-day functioning and are serious public health concerns with social and economic impacts [6].

Although there is growing awareness and support for people with mental health illnesses, 33% of Canadians with mental health needs report that their needs are either not, or only partially, met and that 40% of them prefer to manage their own mental health needs [7]. This is intensified as stigmas attached to mental health problems [8] pose a serious barrier to diagnosis and treatment. Given that individuals with mental health illnesses respond well to early intervention [9], there is hope that early awareness could play a significant role in supporting the mental health needs of Canadians. Early detection of frequent changes in emotions has been recognized as a cornerstone in the treatment of both mood and anxiety disorders.

To this end, researchers in the field of affective computing have been developing systems to elicit and recognize emotion using various modalities. For example, Mower et al. determined emotions from the audio and visual data of actors as they performed emotionally evocative scripts [10]. Valstar and Pantic showed short films to induce disgust, happiness, and surprise in participants [11], and used the Face Action Cod-

ing System (FACS) to determine the experienced emotions. Niu et al. used songs to induce emotions in participants while collecting physiological signals (electrocardiography (ECG), galvanic skin response (GSR), electromyography (EMG), respiration (RESP)) [12]. Likewise, He et al. collected physiological signals but induced emotions using video clips [13]. The authors noted the benefit of physiological signals (ECG, RESP) for individuals experiencing mental illnesses, as these individuals may avoid exhibiting changes in facial expressions, tone of voice, body posture, and gestures. Since emotion is a highly cognitive process, it is particularly valuable to understand how electroencephalography (EEG) signals relate to emotion. Several studies have used EEG for emotion recognition as its measurement uses relatively inexpensive, noninvasive technology that yields high temporal resolution [14, 15, 16]. Due to the increased demand for Human-Computer-Interaction (HCI), the desire to understand emotions, and thus the field of affective computing, has been growing in recent years.

Despite this growth, it can be argued that the field remains limited in applicability due to the experimental conditions employed in most research. The majority of emotion recognition studies are conducted using stimuli that last less than ten minutes [17, 18]. Generally, studies are based on short stimuli, as psychologists have recommended eliciting discrete emotions to alleviate challenges with obtaining ground truth labels for the data [19, 20]. The use of such short stimulus presentation to participants, however, cannot then reflect the temporal and contextual evolution and continuum of emotions. This can be confounded further by the label generation. For example, it is common practice in existing studies to have participants provide an annotation (i.e., happy or sad) at the end of the period of stimulation. This is necessary as having them continuously record annotations throughout the stimulus would preclude immersion for the subject. Providing retrospective annotations, however, is highly error-prone and so participants are commonly asked to provide only a single label for the entire period. This only further adds to the challenge of understanding

the evolution of emotion over time. In contrast, longer stimuli may provide a more dynamic range and context of emotion, facilitating a better understanding of emotion, but provide challenges in acquiring accurate labels.

To facilitate emotion recognition, researchers have used a variety of classification schemes such as Support Vector Machines (SVM) [21, 22] and Linear Discriminant Analysis (LDA) [23], and explored temporal frameworks such as Hidden Markov Models (HMM) [24]. Recently, deep sequence learning techniques, such as long short-term memory (LSTM), have been used to model emotion recognition problems. These techniques may extract meaningful, generalizable, and abstract features, but can be seen more "black box" than conventional classifiers because feature representations are not yet well understood by the research community. Indeed, the design and understanding of appropriate features are a critical aspect of machine learning [25], and can, therefore, provide meaningful insights into how systems work.

1.2 Objective

The long-term motivation of this work is to improve the performance of EEG-based emotion recognition systems to enable better affective computing interfaces and potentially earlier detection of mental health illnesses. The objective of this initial work is to investigate how the consideration of different design factors impacts the performance of EEG-based emotion recognition systems. Focus is placed on performance within the context of dynamically evolving emotion using movies from the AMIGOS dataset [3], with the following specific aims:

Specific Aim 1: To evaluate the body of EEG features found in the literature to inform the design of an emotion-recognition system.

Specific Aim 2: To understand the impact of various cross-validation assessment techniques, and thereby use cases, on classification performance.

Specific Aim 3: To investigate the impact of imperfect labelling and training set selection on emotion classification performance.

1.3 Contribution

Emotion recognition is typically performed using short stimuli for a specific cross-validation scheme, such as leave-one-person-out (LOPO). Cross-validation schemes like LOPO (which are applicable only for a known stimulus) provide limited insight into how well a laboratory-created model may generalize to the real-world.

This work uses the AMIGOS database [3], which includes longer films as stimuli and contains continuous annotations. This provision enables emotions to be analyzed as they evolve and, thus, for new insights to be made, through feature engineering, about feature sets that may be more appropriate for real-world applications.

Furthermore, four cross-validation techniques are evaluated to better evaluate the robustness of the classification models and draw conclusions about the generalizability of selected features. Because model performance is highly influenced by both the training data and training labels, the robustness of training data is evaluated. Techniques such as data selection, gender-disaggregation, and rejection are demonstrated to provide important insights and improvements for the classification problem.

1.4 Thesis Structure

Chapter 2 presents a theoretical background to emotion recognition systems, as well as a review of the state-of-the-art in the literature. Chapter 3 provides an overview and baseline investigation of the AMIGOS database and explores several extensions on their original results using various evaluation frameworks. Chapter 4 presents an investigation of labels and training strategies aimed at improving these baseline results. Finally, Chapter 5 concludes this work and provides ideas for future work.

Chapter 2

Background & Literature Review

2.1 Emotion

2.1.1 Elicitation

Affective computing studies typically adopt one of two strategies to elicit emotion; either they ask the user to self-initiate the emotion or they are exposed to a known stimulus [26, 18, 17].

In self-initiated methods, researchers attempt to elicit emotions by requesting that the participant actively and intentionally evoke an emotion [27]. Participants may, for example, be asked to recall a memory, such as a time when they felt angry [28, 29, 30], or act as they would when experiencing a certain emotion, such as smiling when happy [31, 32].

Alternatively, in exposure-type methods, participants may be asked to partake in a pre-designed event. A common event used to induce known emotions is to have the participant watch a video [33, 34], such as a *Mr. Bean* film to induce amusement and maybe even laughter. Other studies use emotional responses to images [35, 36, 37], with curated image databases allowing standardization across such research [35]. Similarly, participants may either listen to pre-selected music [38, 39, 40], or play video games [23, 41].

Although both videos and audiovisual clips are widely used in emotion-recognition, videos have emerged as a preferred elicitation technique in emotion experiments [17, 18, 42, 43]. Because they are more immersive, they have been found to result in

stronger corresponding physiological changes [20, 42].

2.1.2 Quantification

Emotion is commonly modelled using one of two approaches: a categorical (discrete) model or a dimensional (continuous) model.

The Ekman model, the model most often used for discrete emotions, suggests that there are basic emotions that are not only separate and discrete emotional states, but are also universal [44]. From his research on facial expression, it was determined that there were six basic emotions: anger, fear, sadness, enjoyment, disgust, and surprise [44].

In contrast, the dimensional (continuous) approach allows for systematic inter-relations between emotions. One embodiment of the dimensional approach is Russell's Circumflex Model of Affect [45], which is commonly used to quantify emotions. This 2D model allows for emotions to be represented by two dimensions: valence and arousal. Valence quantifies the range of positive and negative emotion, which ranges from pleasant to unpleasant. Arousal quantifies the level of engagement from passive to active, indicating the intensity of affect. An example mapping of categorical emotions to the valence and arousal dimensions can be seen in Figure 2.1.

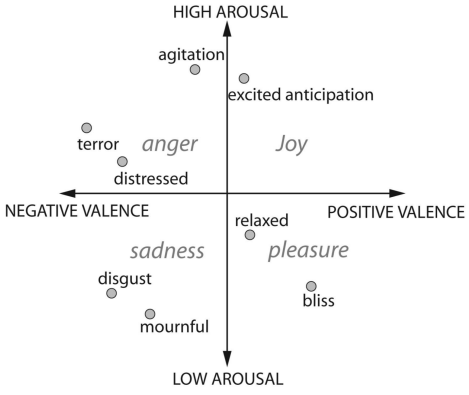


Figure 2.1: Valence-Arousal Model [1]

In a review of current challenges in affective computing [42], Bota et al. discussed the advantages of using this multi-dimensional approach to overcome the difficulties of encoding complex human emotion into words. The work presented in this thesis follows this use of multi-dimensional quantification, but as with much of the machine learning literature, separates each dimension into low and high affect [46, 47, 48].

2.1.2.1 Labelling

Labels (known as the ground truth) created based on these emotion models are used as a representation of an individual's emotion. In affective computing research, particularly that which is focused on machine learning, these labels are assumed to be the true emotional response felt by someone while experiencing a stimulus. This estimation of ground truth is typically determined either from the reported subjective feelings of the subjects themselves or through some form of external examination. Several strategies within these two constructs are used to quantify emotion within an affect model. For example, someone watching a participant that is laughing loudly and slapping their knee may quantify (label) this event as high valence and high arousal using Russell's Circumflex Model of Affect (Figure 2.1).

In subjective reporting, participants may provide retrospective labels about how they felt as they were experiencing the stimulus [49, 50]. This technique is a good indicator of the average emotion felt during the stimulus [20] but is not typically able to capture continuous labels across time. To know what is felt in the moment, the participant could annotate the stimulus as they experience it, but this is widely accepted to impact the participant's emotional response due to the need for in-situ self-assessment [51].

In external examination or labelling, an external annotator may label the stimulus itself or attempt to evaluate the emotion of the subject. When labelling the stimulus, it is assumed that the subjects and annotators experience a similar emo-

tional response. This method, therefore, fails to consider any idiosyncrasies of a subject's mood, history, or personal response [52]. Consequently, many studies have opted to develop labels based on visual cues exhibited by the subject, such as their facial expressions. While this method may provide more customization, it is substantially more onerous (requiring labelling of each subject, as opposed to each stimulus) and inherently relies on visual 'tells'. For this reason, there is no guarantee that the observed response is genuine, particularly when subjects are evaluated in group settings [34].

2.2 Modalities

Even in casual social interaction, physical cues are often interpreted as indicators of emotions. When an individual is excited, for example, it is common for their heart to race, or when nervous, one's hands might become clammy or they may begin to sweat. Similarly, in affective computing, physiological signals serve as inputs to emotion recognition systems. Many researchers have explored the development of robust emotion recognition systems using peripheral signals such as galvanic skin response (GSR), electrocardiography (ECG), electromyography (EMG), respiration (RESP), skin temperature (TEMP), blood volume pulse (BVP), electrooculography (EOG), eye gaze, etc. Information from the central nervous system can be collected by means of electroencephalography (EEG), magnetoencephalography (MEG), and functional near-infrared spectroscopy (fNIRS) and external information can come from audio and video recordings. Of the many available sensors to capture emotion, GSR, ECG, RESP, EEG, and EMG are the most commonly used [42].

Whereas peripheral signals measure the peripheral embodiment of emotion, the neuroscience and affective computing fields often use EEG as a more direct way of interpreting emotion. EEG records a measure of brain activity via electrodes placed

on the scalp. This electrical activity stems from currents that flow during the synaptic excitation of neurons in the cerebral cortex as they are activated [42]. This is important in emotion recognition because the brain's hemispheres have been associated with several mental tasks and states. The left hemisphere is thought to be more involved with language, while the right hemisphere is more involved in nonverbal memory and music, and has special capacities for facial recognition [53]. There are also four lobes that have different functions: the occipital, parietal, temporal, and frontal. The occipital lobe's main function is vision [53], whereas the parietal lobe processes and integrates somatosensory (sensations from the body) and visual information, especially to control movement [53]. The temporal lobe's main functions are to process auditory information, recognize visual objects, emotion, and memory [53]. The frontal lobe's main functions are to determine behaviours, make decisions related to emotion and reward, short-term memory, and attention [53].

Using these electrodes, EEG recordings are minimally-invasive, inexpensive, and inclusive (e.g., participants with limited or no communication abilities can have their emotions captured) [54]. They also provide continuous information with high time resolution. As EEG signals are commonly available, and a center for emotional information, this work will focus on the EEG modality.

2.2.1 Multimodal Systems

Fusing various modalities into a multi-modal system can provide performance benefits as compared to unimodal approaches. It has been suggested that multimodal systems better reflect the complex nature of emotion because different signals may provide complementary information about the various aspects of a participant's emotional response [55]. For instance, a racing heart may provide strong information about arousal, but EEG may be needed to differentiate between excitement or fear.

EEG signals are frequently fused with peripheral and external signals to create

multimodal systems [56, 57, 58, 59, 60]. The fusion of these signals has been shown to improve the classification performance over unimodal EEG systems [60, 61, 62, 63, 64]. Although multimodal systems may yield higher performance than unimodal systems alone, the constituent unimodal systems form their foundation. It is, therefore, important that unimodal systems continue to be explored to enable a clear and targeted understanding of specific signals and systems.

2.2.2 EEG-Based Emotion Recognition

Many researchers have explored the development of robust emotion recognition systems using single (unimodal) modalities. This work focuses on EEG for reasons outlined in Section 2.2, and because it has been shown to outperform other peripheral signals for emotion recognition [28, 65, 56, 60, 63, 66, 58, 67, 68, 69], possibly due to the substantial role the brain plays in regulating and processing sensory inputs and emotion [70].

Various techniques have been investigated to improve emotion recognition performance with EEG, including extracting different features to increase the information density in the signal. Most commonly, features are extracted from specific frequency bands in the signal (e.g., the alpha, beta, delta, gamma, and theta bands are used in 89.4% of works) [18]. To extract these features, frequency (Freq) domain techniques such as power spectral density (PSD) [71, 72, 73] and asymmetry (Asymm) [74, 75] between electrodes have been proposed. Time-domain features can also be extracted, such as high order crossings [76, 77, 78] and Hjorth parameters [79, 80]. Non-linear methods can be used to extract various entropy measures [81, 82] and fractal dimensions [81, 83], and more recently, features have been learned automatically using deep learning approaches [69, 84]. Jenke et al. [75] reviewed various features extracted from EEG signals, and determined that complexity measures such as fractal dimensions are important for emotion recognition, but suggested that further investigation

is needed to better understand the impact of groups of features.

Various classifiers have also been explored to improve emotion recognition performance. Many classical classifiers have been proposed, including support vector machine (SVM) [85, 86, 87, 62, 47, 88, 89, 78, 90, 91], linear discriminant analysis (LDA) [86, 62, 47], k-nearest neighbours (kNN) [86, 87, 92], random forest (RF) [80, 92, 78], Naïve Bayes (NB) [85, 92, 48], extreme gradient boosting (XGB) [85], and decision trees [86, 87, 92]. The performances achieved by the classifiers vary, but differences are often overshadowed by the impact of the features chosen as part of these models. In recent deep learning works, convolutional neural networks (CNN) [47, 93], recurrent neural networks (RNN) [93], long short-term memory (LSTM) networks [84, 94], and extreme learning machines (ELM) [69] have all been employed. Motivated by these works, this thesis focuses on the development and improvement of EEG-based emotion recognition.

2.3 Emotion Databases

Because of growing interest in the field, and because many research groups cannot collect physiological signals for affective computing, several databases have been made publicly available. This section reviews many of these datasets and provides a taxonomy based on the signals collected, the labelling methods, and the stimuli used. Table 2.1 summarizes an overview of the findings, whereas the remainder of this section presents a brief discussion of the datasets. The legend for the values of the labelling strategies in Table 2.1 is as follows:

Self1 - A single self-assessed annotation per elicitation (e.g., a single user-generated label per movie)

Ext1 - A single externally-assessed annotation per elicitation (e.g., a single label generated per movie by an external observer)

SelfCont - A sequence of retrospective self-assessments (e.g., one user-generated label every twenty seconds of a movie)

ExtCont - A sequence of externally-assessed annotations (e.g., a label generated every twenty seconds per movie by an external observer)

SelfAct - Self-elicited emotion (e.g., sadness acted out by the participant based on a sad memory)

Table 2.1: Overview of Affective Databases Using Physiological Signals

Database	**Modalities**	**Subject Count**	**Stimulus & Length**	**Label Strategy**	**Valence Arousal**
AMIGOS 2018, [3]	Video, ECG, GSR, EEG	40	20 clips 1-24 min	Self1 ExtCont	Yes
ASCERTAIN 2018, [64]	GSR, EEG, ECG, Facial Landmark Trajectories	58	36 clips 1-2 min	Self1	Yes
BioVid Emo 2016, [95]	GSR, ECG, EMG	94	15 clips 0.5-4 min	Self1	Yes
DEAP 2011, [48]	EEG, GSR, BVP, EMG, EOG, RESP, Video, TEMP	32	40 clips 1 min	Self1	Yes
DECAF 2015, [96]	MEG, ECG, EMG, EOG, Video	30	76 clips 1-2 min	Self1	Yes
DREAMER 2018 [97]	EEG, ECG	25	18 clips 1-6 min	Self1	Yes
EMDB 2012, [98]	GSR, HR	32	52 clips 0.7 min	Ext1	Yes
Gjoreski et al. 2017, [99]	BVP, GSR, TEMP	21	Stressful Tasks 17-40 min	Self1	No
Healey et al. 2005, [100]	ECG, EMG, GSR, RESP, Video	9	Driving 50 min	Self1 ExtCont	No
Liu et al. 2018, [33]	EEG	30	16 clips 1-2 min	Self1	Yes
MAHNOB-HCI 2012, [67]	EEG, GSR, ECG, Eye Gaze, Video, Audio, TEMP	27	20 clips 0.5-2 min	Self1	Yes
Savran et al. 2006, [101]	EEG, RESP, BVP, GSR, fNIRS	5	327 Images 2.5 s	Self1	Yes

Database	Modalities	Subject Count	Stimulus & Length	Label Strategy	Valence Arousal
Schneegass et al. 2013, [102]	GSR, ECG, TEMP, Context Data	10	Driving 30 min	SelfCont	No
SEED 2015, [103]	EEG, Video	15	15 clips 4 min	Self1	No
SEED-IV 2019, [104]	EEG, Eye Tracking	15	72 clips 2 min	Self1	No
SEED-VIG 2017, [105]	EEG, EOG	23	Driving 120 min	ExtCont	No
SWELL-KW 2014, [106]	ECG, GSR Video, Audio, Computer Logging	25	Stressful Tasks 30-45 min	Self1 ExtCont	Yes
Vyzas et al. 1999, [107]	EMG, BVP, GSR, Respiration	1	Acting 3 min	SelfAct	No
WESAD 2018, [108]	EMG, ECG, GSR, BVP, TEMP, Motion, RESP	15	11 clips, Stress Test [109] 0.5-10 min	Self1	Yes
Yazdani et al. 2012, [68]	EEG, GSR, EMG, BVP, RESP, EOG, TEMP	6	20 clips 2 min	Self1	Yes

Due to the challenges associated with labelling emotions, most studies have limited experiments to short stimuli, during which emotion has been assumed to be static, or stationary. To elicit only one emotion, many research groups have used stimulus lengths of less than 10 minutes [17, 18].

Conversely, longer and more dynamic stimuli can elicit more than one emotion, resulting in transitions between emotions. These long stimuli motivate the exploration of the temporal dynamics of emotion in affective computing but raise substantial challenges with labelling. This is evidenced in Table 2.1 by the lack of datasets that contain sequences of labels (i.e., SelfCont or ExtCont).

Although some research has used continuous labelling, most of these have focused on computer vision applications and do not include physiological signals [110, 111, 112]. In 2016, Soleymani et al. [94] recorded EEG for short movies, and external annotators were asked to label frontal-view videos of participants watching the films.

More recently, the AMIGOS [3] dataset was released, comprising ECG, EEG and GSR data from both short movies (less than 150 seconds) and longer movies (greater than 14 minutes). Most importantly, this dataset includes regular emotion labels as determined by three external annotators at twenty-second intervals. These ground truth labels were generated by watching the facial expressions of participants and were quantified using the valence-arousal scale [45]. The recent availability of AMIGOS's continuous set of labels for longer movie sequences presents an opportunity to explore the performance of emotion recognition systems when emotion is not assumed to be static. Consequently, the remainder of this work focuses on the use and design of EEG-based emotion classification using the AMIGOS database.

2.4 Evaluation Frameworks

In addition to the various parameters already discussed in emotion recognition, the framework used to evaluate system performance can have a substantial impact on the results and their interpretation. In a review of emotion recognition systems using physiological signals, Shu et al. [17] reported that most existing systems were validated using k-fold or leave-one-sample-out (LOO) techniques. These common validation frameworks (which use some portion of the data for training while reserving some for testing), however, are agnostic of the application and; therefore, further consideration of the problem is needed.

In affective computing, there are two main frameworks most often used when developing physiological models for emotion recognition. The first method, the participant agnostic (independent) approach, is used in the largest proportion of the literature [48, 85, 66, 113, 114]. Because it is designed to extend to previously unseen participants, the generalizability of this approach could make it more tangible for real-world adoption. Such participant-independent systems, however, have often obtained

F1-scores (harmonic mean of precision and recall, further explained in Section 3.3.2) between 0.5 to 0.6 [48, 96, 59].

The second framework focuses on developing a model for a given participant (participant-dependent or within-subject). The participant-dependent approach has yielded substantially better performance [17, 115]; typical F1-scores fall between 0.6 to 0.8 [116, 117]. Although within-subject models produce better results, their real-world applicability may be limited because they require a model to be trained specifically for each user. Nevertheless, this framework offers the ability to evaluate the information content without the confounding factor of inter-subject variability. This is particularly important for EEG [118]. Although it is generally found to be a better predictor of emotion than other physiological signals [48, 3, 67], EEG varies greatly between participants [118, 119] and can be heavily influenced by group-dynamics [120, 121].

While k-fold and LOO approaches are popular in literature, it has been suggested that for participant-independent results, frameworks such as Leave-One-Person-Out (LOPO) be applied [42]. In this framework, the system is trained with all subjects but one and tested with the previously unseen subject. This process is repeated until all subjects have been tested, and the results are averaged across all cases.

Within the participant-dependent approach, the k-fold approach has been more prevalent [17] although some works have also used LOO [117, 122]. Depending on their application, however, these approaches may also provide the classifiers with training information about a known stimulus. For example, parts of the same movie could be selected for training and testing, resulting in an emotion recognition system that is dependent on having previously seen a specific stimulus. This further constrains the generalizability of such a system to the point where tangible applications may be limited. Consequently, in this work, any discussion of participant-dependent results are based on a stimulus-independent Leave-One-Movie-Out (LOMO) framework, meaning that all movies but one are used for training, and testing is performed

on the remaining, previously unseen, movie. This process is repeated until all movies have been tested, and the results are averaged across all cases. While not as popular as LOPO, the LOMO framework has recently been explored in the literature. Malandrakis et al. [110] and Baveye et al. [111] used LOMO with audio and video features to classify affective states. Similarly, Tian et al. [123] used the LOMO scheme to investigate how well audio, video, and GSR features can determine the level of induced valence and arousal in an audience.

One important consideration with the LOPO approach is that it is most often applied for a given stimulus. For example, although the testing subject may be previously unknown, the data for all subjects are collected while they watch the same movie. Again, this conceptually limits the generalizability of the results to new people, but only during known stimuli. A truly subject and stimulus-independent Leave-One-Person-And-Movie-Out (LOPMO) scheme has not yet been widely adopted in the literature, presumably because it severely reduces the observed performance of emotion recognition systems. Nevertheless, LOPMO best reflects the goal of a real-world implementation, as it would represent a truly generalizable subject- and stimulus-independent affective computing system. Indeed, even in 2011, Kolodyazhniy et al. [124] suggested the need for a subject and stimulus-independent classification. In their work, they used ten-minute movies that elicited fear, sadness, and neutral emotional states, and a total of 14 features derived from ECG, GSR, respiration, temperature, and EMG. Their work, however, did not employ EEG or use continuous labels.

In summary, these evaluation frameworks can be categorized as shown in Table 2.2. Note that the real-world applicability of these frameworks increases from a minimum in the top left (known subject and stimulus), to a maximum in the bottom right (LOPMO). In this thesis, results are presented for the LOPO, LOMO, and LOPMO cross-validation evaluation frameworks.

Table 2.2: Cross-validation Evaluation Frameworks

	Known Subject	**Unknown Subject**
Known Stimulus:	Subject- and stimulus-dependent (k-fold, LOO)	Subject-independent (LOPO)
Unknown Stimulus:	Stimulus-independent (LOMO)	Subject- and Stimulus-independent (LOPMO)

2.5 AMIGOS Research

Because of its unique combination of longer movies, multi-modality, and continuous labels, several research groups have used the AMIGOS dataset since its release in 2018. A summary of these works is presented in Table 2.3 along with which modalities they used, whether they used the long or short movies (many groups use only the short movies because of the assumed stationarity of emotion), the evaluation technique, and details about the classifier(s) used.

Table 2.3: AMIGOS Literature Review

Paper	**Modality**	**Length**	**Classifier(s)**	**Evaluation**
Miranda-Correa et al. 2018, [93]	EEG	Long & Short	CNN, RNN, Fusion	LOPO
Siddharth et al. 2018, [66]	EEG, ECG GSR	Short	ELM	10-fold
Siddharth et al. 2019, [69]	EEG, ECG GSR	Short	ELM	LOPO
Wang et al. 2018, [125]	EEG, ECG GSR	Short	XGB, NB, SVM	LOPO
Yang et al. 2019, [126]	EEG, ECG GSR	Short	SVM	LOPO
Tung et al. 2019, [113]	EEG, ECG GSR	Short	NB, XGB	LOPO
Shukla et al. 2019, [116]	GSR	Long & Short	SVM	Train-Test-Split
Yang et al. 2019, [127]	EEG, ECG GSR	Long & Short	Linear SVM	10-fold

Paper	Modality	Length	Classifier(s)	Evaluation
Li et al. 2020, [128]	EEG, ECG GSR	Long & Short	NB, CNN, LSTM	LOPO
Mou et al. 2019, [117]	Video	Long	Linear SVM, SVR, LSTM	LOPO, LOO (Within), LOGO
Santamaria-Granados et al. 2019, [129]	ECG & GSR	Short	NB, kNN, LDA, RF, Linear SVM, Multi-layer Perceptron, AdaBoost, DCNN	Train-Test-Split
Zhao et al. 2019, [63]	EEG, ECG GSR	Short	3D CNN, 1D CNN	10-fold
Harper et al. 2019, [130]	ECG	Short	LSTM	LOPO
Harper et al. 2019, [131]	ECG	Short	LSTM	LOPO
Chang et al. 2019, [85]	EEG, ECG GSR	Short	NB, SVM, XGB, Hyperdimensional Computing	LOPO
Gjoreski, et al. 2018, [114]	ECG & GSR	Short	XGB, SVM, kNN, RF, AdaBoost, NB, Decision Trees	Movie-specific 10-fold
Sarkar et al. 2019, [132]	ECG	Long & Short	CNN	10-fold

Of the works in Table 2.3, only six have presented results for unimodal EEG data. Some groups have used the EEG information as part of multimodal systems with the other modalities. It can also be seen that while the classifiers used to create the models vary, there is a strong trend towards using LOPO cross-validation to evaluate model performance. Table 2.4 presents a summary of the classification performance (using F1-score) of the systems developed using EEG, along with the results from the original AMIGOS work as the baseline. Note that some works used EEG but did not report F1-score, which is an important metric in classification when classes are unbalanced [66, 126]. It has been noted that the AMIGOS dataset is highly imbalanced, with different numbers of exemplars from each class (high/low valence and arousal), as is further discussed in Chapter 4.

As can be seen in Table 2.4, most F1-scores for the AMIGOS database have been between 0.550 and 0.600, indicating the difficulty of the classification problem. It should also be noted that a majority of AMIGOS-based research results have only leveraged the short films. Two research groups achieved F1-scores greater than 0.600. Tung et al. [113] used the extreme gradient boosting (XGB) classifier and recorded 0.753 valence and 0.568 arousal F1-scores using PSD, asymmetry, and entropy-domain features. Siddharth et al. [69] fused traditional features with those obtained from deep learning, resulting in F1-scores of 0.800 and 0.740 for valence and arousal. While deep learning features improve emotion recognition, their understanding remains limited, and so this work focuses on how traditionally engineered features relate to emotion.

Table 2.4: Review of Works Using F1-score to Evaluate Unimodal EEG

Paper	**Classifier**	**Feats**	**CV**	**Length**	**F1-score**	
					Valence	**Arousal**
AMIGOS 2018, [3]	NB	PSD & Asymm	LOPO	Long	0.557	0.571
				Short	0.576	0.592
Siddharth et al. 2019, [69]	ELM	PSD, Deep & Entropy	LOPO	Short	0.800	0.740
Miranda-Correa and Patras 2018, [93]	CNN	PSD & Time	LOPO	Short & Long	0.580	0.570
	RNN				0.570	0.590
	Fusion				0.590	0.610
Wang et al. 2018, [125]	XGB	PSD & Asymm	LOPO	Short	0.577	0.604
	SVM				0.556	0.557
Tung et al. 2019, [113]	XGB (1)	Time, Freq & Entropy	LOPO	Short	0.575	0.568
	XGB (2)				0.753	0.568

Chapter 3

Classifying Emotion from EEG Signals

As outlined in the previous chapter, EEG data from the AMIGOS database were used in this work. To establish the performance of the developed system and to compare with established norms, a conventional pattern recognition framework was followed. Per this approach, preprocessed EEG signals from the AMIGOS dataset were adopted for windowing, feature extraction, training, and classification.

3.1 Database Description & Preprocessing

The AMIGOS [3] dataset contains EEG data from 40 participants, recorded while they watched 16 different short films (<150 seconds each) and excerpts from 4 longer films (>14 minutes each). The short videos were watched individually in isolation, whereas some participants watched the long films together in groups. The long films consisted of excerpts from the following four movies:

- *The Descent*, a 2005 horror thriller about a caving expedition gone wrong [1]
- *Back to School, Mr. Bean*, an episode of the 1990 comedy television show [2]
- *The Dark Knight*, a 2008 fantasy crime drama about 'The Joker' [3]
- *Up*, a 2009 animated movie about a 'feel good' family adventure [4]

[1] https://www.imdb.com/title/tt0435625/
[2] https://www.imdb.com/title/tt0651839/
[3] https://www.imdb.com/title/tt0468569/
[4] https://www.imdb.com/title/tt1049413/

The EEG signals were collected using an Emotiv EPOC Neuroheadset (14 channel, 128 Hz sampling rate, 14-bit resolution) [133]. The publicly available dataset, AMIGOS [3], includes both the raw EEG signals as well as preprocessed versions for all participants. The preprocessed signals, used in this work, were average-referenced and high-pass filtered with a 2 Hz cut-off frequency. Motion artifacts from eye movements were removed using a blind source separation technique [134].

The ground truth for each participant and film is also provided with the AMIGOS dataset, which is critical for building and validating the model. The ground truth was determined in two ways. The first method was to have participants self-report their level of valence, arousal, dominance, liking, and familiarity, and their experienced emotion (selected from a list of basic emotions that include: Neutral, Disgust, Happiness, Surprise, Anger, Fear or Sadness). For the long movies, this self-assessment was only completed after two movies had been watched. The second approach was to have external annotators examine the facial expressions of the participants, which were video-recorded along with the physiological signals as the participants watched the films. These external labels were created for each twenty-second period. The first twenty-second clip began five seconds before the start of the video and overlapped with the second twenty-second label, which started from zero. All subsequent labels corresponded to non-overlapping segments of twenty seconds beginning at twenty seconds and continuing until the end of the film. Instead of shortening the duration of the final clip, the last label covered the period of twenty seconds from the end of the film, creating an overlap with the second-to-last clip.

These clips of the facial expressions were rated by three external annotators using the continuous valence and arousal scales. Before annotation, the order of the twenty-second clips was randomized for each participant, but the annotators viewed these sequences in the same order. Each annotator's ratings for every twenty-second clip of each movie are provided as part of the AMIGOS dataset.

Participants 8, 24, and 28 were missing the external annotations for all movies and were, therefore, not used in this work. Participants 17, 18, and 22 were excluded from any LOMO-Within and LOPMO evaluations as they were missing external annotations for some long movies. For the LOMO-Within case, participants were only included if they had at least ten high affect samples within all training set combinations. This ten sample threshold represents approximately a 6% high affect class and was chosen to balance the number of testable models and the class imbalance. The excluded participants are [20, 21, 23, 26, 31, and 38-40] for valence and [11-13, 15, 20, 21, 23, 25-27, 29, 30, 35, 37, and 40] for arousal. Please refer to Section 4.2 for a more detailed analysis and discussion about class imbalance.

3.2 Feature Extraction

The information content of the 14 channels of EEG from the Emotiv headset is related to the position of the electrode on the head. Per the 10-20 system (an internationally recognized method of describing electrode locations on the scalp) [135], starting at the front-left side of the head and proceeding in a counter-clockwise direction, the Emotiv channels correspond to AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. As shown in Figure 3.1, the odd-numbered electrodes are on the left side of the head, and the even numbers are on the right. The letters refer to the placement of the electrodes with regards to the lobes. 'O' indicates the occipital lobe, 'P', the parietal lobe, 'F', the frontal lobe, and 'T', the temporal lobe. 'AF' refers to the anterior part of the frontal lobe. 'FC' refers to the frontal lobe, but towards the center.

To avoid having two overlapping frames of data as outliers, the first and last labels for each movie were discarded. This left a set of 14 different time-series of non-overlapping twenty-second windows of EEG from which features were extracted.

The original results published along with the AMIGOS database [3] included the use of 105 EEG features. These included the power spectral density (PSD) of five bands of EEG for each electrode: theta (3-7 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (14-29 Hz), and gamma (30-47 Hz) (5 PSD from bands x 14 channels = 70 features). The differential spectral power asymmetries between the seven left/right hemispheric pairs of electrodes across the five bands were also extracted (5 bands x 7 channel asymmetries = 35 features). The electrode pairs can be seen in Figure 3.1, such as T7 and T8.

To compare with and build upon these results, the aforementioned 105 features were extracted along with an additional 112 identified via literature review and compiled from various sources (for a total of 217). Briefly, these included the *fractal dimension* (2 methods), *entropy* (2 methods), *Hjorth parameters* (2 methods), *detrended fluctuation analysis*, and *Fisher information*. Features were implemented using Python version 3.8, and the PyEEG [136] and EntroPy [137] toolboxes. The entire set of features is explained in more detail below.

- **PSD**: The power spectral density, reflective of the distribution of signal power across frequencies [138], is one of the most common EEG features. The PSD was calculated for different EEG frequency bands, as different bands have been

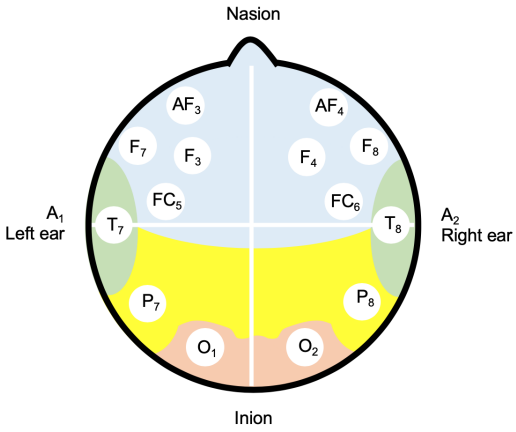


Figure 3.1: 10-20 EEG Placement, Adapted From [2]

associated with different processes [139, 18]. Theta waves are associated with affective processing [140], whereas the slower alpha band reflects attentional demands such as alertness and expectancy [141]. The entire alpha band reflects task-related processes [141]. Alpha waves tend to occur when someone is in a relaxed state of mind, while beta waves tend to occur when an individual is in more of an active state [46]. The gamma band reflects a reflective aspect when processing emotional material [142].

Following AMIGOS [3], the Welch method with windows of 128 samples (1s) were used to calculate the PSDs. These PSDs, X, were then averaged over each frequency band, and the logarithms were obtained as features.

$$PSD_{f_{\text{low}},f_{\text{high}}} = \log\left(\frac{1}{f_{\text{high}} - f_{\text{low}}} \sum_{f=f_{\text{low}}}^{f_{\text{high}}} X[f]\right), \tag{3.1}$$

- **Spectral Asymmetry (Asymm)**: Spectral asymmetry leverages both frequency-domain and spatial information about emotional changes in the brain [73]. The asymmetry in the different bands across channels has been shown to be indicative of different emotions. For example, the alpha asymmetry between frontal lobe channels F3 and F4 can relate to valence [143] and the beta asymmetry between parietal lobe channels P3 and P4 can correlate with angry facial expressions [144].

  The differential spectral asymmetry was calculated by taking the difference of the PSD features from symmetric channels from the left and right hemispheres

$$Asymmetry^{a,b}_{f_{\text{low}},f_{\text{high}}} = PSD^{a}_{f_{\text{low}},f_{\text{high}}} - PSD^{b}_{f_{\text{low}},f_{\text{high}}}, \tag{3.2}$$

  where a and b represent different EEG channels. For example, a slow alpha asymmetry for parietal lobe channel P3 and P4 is $Asymmetry^{\text{P3,P4}}_{8,10}$.

- **Hjorth**: The Hjorth Mobility (HM) is an estimation of the signal's mean frequency, and the Hjorth Complexity (HC) reflects the bandwidth and the change in frequency [145]. It is defined as the square root of the variance of the signal derivative, normalized by the variance of the signal. While not yet widely adopted, Hjorth features have been shown to be relevant for emotion recognition [79]. Equations are as given in [136].

- **Detrended Fluctuation Analysis (DFA)**: DFA quantifies the statistical persistence, or auto-correlation, property of non-stationary physiological signals [146]. Briefly, DFA evaluates the detrended and integrated signal as a function of window size. Commonly used in many fields, including for ECG analysis, DFA has also been found to be beneficial for EEG emotion recognition [147]. Equations are as given in [136].

- **Fractal Dimension (FD)**: Fractal dimension approaches, such as the Petrosian fractal dimension (PFD) and Higuchi fractal dimension (HFD), are a measure of signal complexity [148] and are commonly used for non-stationary and transient signals. The Higuchi fractal dimension has been used more frequently in emotion recognition works [18], but in neurophysiology, both Higuchi and Petrosian fractal dimensions are commonly cited [149]. Equations are as given in [136].

- **Entropy (Ent)**: Entropy is a measure of chaos, or disorder, in a system or signal, and is, therefore, used to understand signal complexity [150]. Here, the spectral entropy (SpecEnt), the entropy of the PSD [137], and the SVD entropy (svdEnt), an indicator of how many vectors are needed to reconstruct an adequate explanation of a signal [137], were used. Hatamikia and Nasrabadi [81] found that spectral entropy outperformed the Petrosian and Katz fractal dimensions for emotion recognition. Gupta et al. [80] used SVD entropy as part

of a set of features to classify discrete emotion for short movies. Equations are as given in [137].

- **Fisher Information (FI)**: The Fisher Information is a measure of how much information a random variable carries about the data that it models. It is also known as the expected value of the observed information [151]. Although less commonly used, FI of EEG has been shown to contain affect information [80]. Equations are as given in [136].

3.2.1 Feature Selection

In machine learning, the selection of features with high information density is essential to improve performance and avoid the 'curse of dimensionality', wherein the training data become sparse due to the high dimensionality introduced by the large number of features. To overcome this problem and identify which features were most relevant and informative for emotion recognition, two common greedy-search feature selection techniques, namely Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), were implemented. The SFS method starts with an empty set of features and incrementally adds a feature to the set. The feature that is added at each step is the one that gives the largest performance improvement when added to the existing set. Conversely, SBS begins with the group of all features and iteratively removes one feature at a time. The feature removed at each step is the one that leaves the highest performance for the remaining set. It should be noted that, although faster than an exhaustive brute-force search, both SFS and SBS (due to their wrapper-style objective functions) may miss elements of information in the relationships between groups of features. Nevertheless, they are both commonly used and, when interpreted carefully, can lead to important understanding and performance benefits.

3.3 Classification

3.3.1 Classifiers

Virtually all classifiers have been used in some form in emotion recognition [18, 17]. The most common classifier, however, is the Support Vector Machine (SVM), reported to have been used by 59% of cases in a recent review [18]. Because of this and its inherent tradeoff between accuracy and generalization [71], SVM was adopted in this work using a linear kernel. Although less frequently used (6.3% of emotion research [18], Linear Discriminant Analysis (LDA) is also robust and less computationally intensive. As a generative classifier, LDA models the distributions of the class data, whereas the discriminative SVM classifier focuses explicitly on the data at the boundaries between classes.

These classifiers were implemented using scikit-learn [152]. The SVM classifier (sklearn: LinearSVC) had the *dual* parameter set to false to ensure convergence for all cross-validation schemes. For LOMO-Within, the *max_iter* parameter was increased from 1000 to 20000 to ensure convergence as well. To accommodate for class imbalances, the learned weights were automatically adjusted to be inversely proportional to the class frequencies in the input data for all CV schemes (the *class_weight* parameter was set to *balanced*). This technique provides more infrequent classes with a higher weight than the more frequent classes. LDA used the default settings provided by the scikit-learn package. Separate classifiers were trained for valence and arousal, as is most commonly done in the field [153, 59].

3.3.2 Evaluation Metrics

Accuracy (as shown in Equation 3.3) is a common metric used to measure the proportion of correctly classified samples but can be misleading for imbalanced classes.

If, for example, one class represents 90% of all samples, a 90% accuracy could be achieved by always predicting that class.

Consequently, in this work, F1-score was used instead to describe the classification performance of the various models. The F1-score is described as the harmonic mean of precision and recall. Precision, as shown in Equation 3.6, determines how many instances of the positive predictions came from the positive class, making it a measure of exactness. Recall, as shown in Equation 3.5, determines how many instances of the positive class are successfully predicted as positive, making it a measure of completeness. F1-score was calculated using the Python sklearn toolbox [152] with the average metric set to *macro*, meaning the metric was calculated for each label, and the mean was taken of these results. As F1-score was calculated with cross-validation, an F1-score was calculated for each partition and then the average F1-score was then reported for across the partitions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

$$\text{F1-score} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{3.5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.6}$$

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative as seen in Table 3.1.

Table 3.1: Description of True and False Positive and Negative Cases

		Predicted	
		Positive	*Negative*
Actual	*Positive*	TP	FN
	Negative	FP	TN

3.4 LOPO Cross-validation

In the original AMIGOS results [3], the evaluation was framed as two two-class problems; high *vs.* low valence, and high *vs.* low arousal. A Naïve Bayes classifier was used to create the model and the LOPO cross-validation evaluation framework was used to validate the model. Again, this framework represents the case where models are trained on a set of users for a given set of stimuli and evaluated on a previously unseen user exposed to the same stimuli. Data for all four long movies were included in the training and testing sets. Data from participants 8, 24, and 28 were unavailable for the long video experiments, thus LOPO was performed using the remaining 37 participants. Feature selection was performed using Fisher's linear discriminant (FLD) [154] to select an unreported number of features. All results were reported using the F1-score metric, achieving an EEG-based performance of 0.557 for valence and 0.571 for arousal for the long movies.

In this work, the performance of both LDA and SVM classifiers was evaluated after completing either SFS or SBS. Feature selection was conducted with all 217 features and the best feature set was determined by selecting the set that achieved the maximum classification performance. This was determined using features extracted from the twenty-second windows of preprocessed EEG signals corresponding to the given labels. For the LOPO scheme, folds were conducted where each of the 37 participants was left out, and the average results across all participant folds were used to select the feature set. Table 3.2 shows the results of these analyses, along with the results of the full feature set, and the original AMIGOS work.

Before testing the statistical significance of the results, a test for normality was conducted using the Kolmogorov-Smirnov test [155] along with a visual normality check using histogram plots to determine if the F1-scores followed a normal distribution. Both tests determined that the distributions were not normal, with an

example histogram plotted in Figure 3.2. Therefore, the non-parametric Kruskal-Wallis H-test [156] was chosen to test for statistical significance. The null hypothesis for this test is that the population medians for the distributions are equal. Significance was calculated between the performance of the feature selection techniques for each classifier and also between the two classifiers. The Python scipy.stats [157] and matplotlib [158] toolboxes were used to implement these techniques.

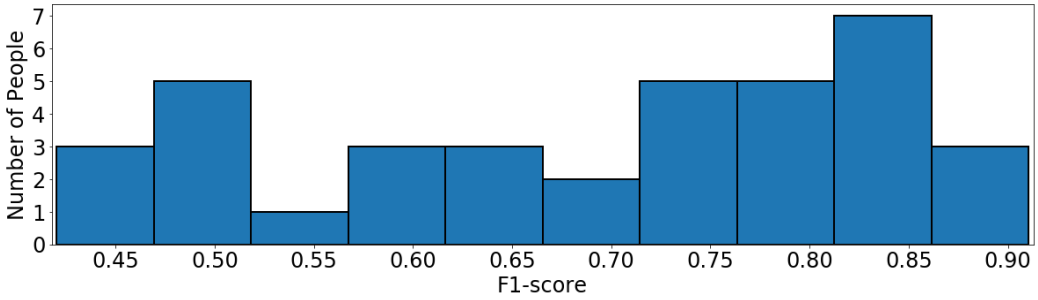


Figure 3.2: Test of Normality for the F1-scores from LOPO, SFS, Valence

As can be seen in Table 3.2, the results obtained here substantially outperform those presented in the original work by Correa et al. [3], although no statistical comparison could be made with the single reported value for each affect. While there was no significant difference between the results of SFS and SBS when using an LDA classifier (arousal $p = 0.585$, valence $p = 0.931$), Figure 3.3 shows how the progression of each approach differs. Because SFS cannot predict correlations between features, it can be seen to be less 'smooth' in its performance improvements. Conversely, SBS starts with all features and removes the worst-performing feature, helping it to better understand feature relationships. These results suggest that there is meaningful information in the correlation between channels and features that plays an important role in EEG-based emotion recognition. Furthermore, the improved performance when using feature selection confirms the presence of the curse of dimensionality when using all 217 features.

Table 3.2: LOPO (Participant-Independent) Results for Each Classifier (CLF) and Feature Selection (FS) Technique

CLF, FS	**Valence**		**Arousal**	
	F1	*#Features*	*F1*	*#Features*
AMIGOS [3]	0.557	-	0.571	-
LDA, SFS	0.687	24	0.683	168
LDA, SBS	0.695	77	**0.699**	63
LDA, ALL	0.648	217	0.664	217
SVM, SFS	0.680	72	0.676	75
SVM, SBS	**0.706***	45	0.694	49
SVM, ALL	0.648	217	0.643	217

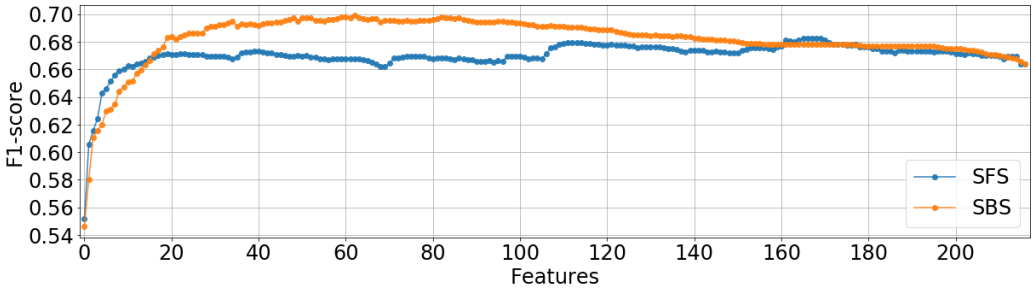
* denotes a significant difference from the CLF, ALL results ($p < 0.05$)

The top five features selected for LOPO cross-validation via SFS and SBS are presented in Table 3.3. Features are named according to the following naming convention: Electrode, Feature, Band. For asymmetry features, the lobe is specified as opposed to the electrode. Only PSD and Asymm features have specified bands. Detailed electrode, feature, and band information can be found in Section 3.2.

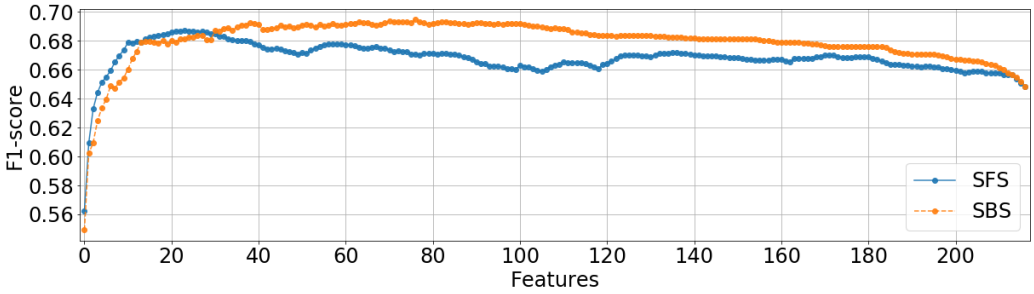
Interestingly, although some of the original AMIGOS features were selected, the majority (and first) of those chosen were previously unused with the AMIGOS dataset. The bottom row of each column shows the F1-score when using these five features, and the percentage of the maximum score these five features achieve. In each case, using only five features achieved over 89% of the best F1-score for each model.

For many of the models, HFD, the Higuchi Fractal Dimension, was found to be the most important feature. This corroborates the findings of previous works, supporting that HFD is an important feature for EEG-based emotion recognition [159, 83].

It can also be noted that SFS and SBS chose different features for LDA. While both found HFD to be the most important feature, a different channel was selected, and the subsequent features were different. As mentioned previously, SBS is better able to understand relationships between features, so perhaps the most highly selected



(a) Arousal Performance for Best Feature Combinations



(b) Valence Performance for Best Feature Combinations

Figure 3.3: Comparison of SFS and SBS Approaches for Classification of Valence and Arousal Using LOPO LDA

features had relationships that SFS was unaware of. Nevertheless, SFS does have a higher performance for both valence and arousal when using only five features. This trend is displayed in Figure 3.3, where SFS increases more rapidly. As LOPO had similar performances between SFS and SBS, perhaps the relationships between the features did not have a significant influence on this classification model.

Table 3.3: Top Five LOPO Features Selected by Feature Selection

		LDA		SVM	
		Valence	*Arousal*	*Valence*	*Arousal*
SFS	1	**T7 HFD**	**T7 HFD**	T7 PSDγ	T7 PSDγ
	2	**AF4 HFD**	**AF4 HM**	AF4 PSDγ	**P8 HFD**
	3	**T8 HFD**	T7 PSDα	**P7 HFD**	**AF3 HFD**
	4	**F8 PFD**	T7 PSDβ	FC Asymmγ	T7 PSDβ
	5	**FC6 HM**	**T8 HFD**	F34 Asymmθ	**F4 SpecEnt**
		0.651 (95%)	0.643 (94%)	0.644 (95%)	0.632 (94%)
SBS	1	**T8 HFD**	**T8 HFD**	**T8 HFD**	**T7 HFD**
	2	**AF4 HFD**	**AF4 HFD**	FC5 PSDθ	**F7 svdEnt**
	3	**F7 PFD**	T7 PSDγ	FC5 PSDγ	FC5 PSDα
	4	**FC5 HM**	T7 PSDβ	F7 PSDγ	**F7 HFD**
	5	**O2 HC**	**F8 HM**	**O2 HC**	FC5 PSDβ
		0.633 (91%)	0.620 (89%)	0.639 (91%)	0.629 (91%)

*__bold features__ were not used in the original AMIGOS set

3.5 LOMO-Inter Cross-validation

For the LOMO-Inter cross-validation evaluation framework, the classification results, along with the number of features selected using SFS and SBS, are shown in Table 3.4. This framework evaluates how well a model trained across multiple users can extrapolate emotional information about those users to a new movie. The data for three movies from all 37 people (8, 24, 28 excluded, as in LOPO) were used to train the model, while the data for these individuals watching the remaining movie was used to test the model. The classification performance for each feature selection iteration was determined by averaging the results across the left-out movies. Note that the original AMIGOS work [3] did not evaluate the LOMO-Inter framework, although it has been explored in many other works for different datasets. When the LOMO-Inter results are compared to the LOPO results, a decrease in performance is noted for both the LDA and SVM models. This decrease indicates that inter-movie variability may have a greater effect on emotion recognition than inter-subject variability. As these results

yielded only four F1-scores from which to estimate the LOMO-Inter distribution, a test for normality was not performed for this scheme. Statistical significance was still calculated using the Kruskal-Wallis H-test; however, the results lack statistical power for this framework.

Table 3.4: LOMO-Inter (Participant-Independent) Results for Each Classifier (CLF) and Feature Selection (FS) Technique

CLF, FS	**Valence**		**Arousal**	
	F1	*#Features*	*F1*	*#Features*
LDA, SFS	0.636	119	0.675	99
LDA, SBS	0.646	68	**0.684**	63
LDA, ALL	0.604	217	0.647	217
SVM, SFS	0.632	115	0.650	131
SVM, SBS	**0.648**	73	0.657	67
SVM, ALL	0.610	217	0.626	217

No statistically significant differences ($p > 0.05$) were found

For both LDA and SVM, SBS trended towards outperforming SFS, but not significantly. Similarly, no significant differences were found between LDA and SVM. Each model also chose a different number of features, with SBS choosing fewer for both classifiers. Table 3.5 shows the top five features selected for the LOMO-Inter framework by SFS and SBS. From these tables, it is again evident that the new features extracted in this work play a significant role in the classification performance. Using five features, over 87% of the maximum F1-score was achieved in all cases. It is also noteworthy that both the LDA and the SVM selected several new features introduced here. As was observed for LOPO, the fractal dimension features were again highlighted as adding valuable information for the emotion recognition problem.

Table 3.5: Top Five LOMO-Inter Features Selected by Feature Selection

		LDA		SVM	
		Valence	*Arousal*	*Valence*	*Arousal*
SFS	1	T7 PSDγ	T8 PSDγ	T7 PSDγ	T8 PSDγ
	2	AF3 PSDβ	**T7 HFD**	AF4 PSDβ	**T7 HFD**
	3	**F8 HFD**	**F7 HM**	**T8 HFD**	**AF4 HC**
	4	**AF4 HC**	**T7 svdEnt**	**AF4 HFD**	**T7 svdEnt**
	5	**T8 HFD**	FC5 PSDβ	**F8 HFD**	P Asymmγ
		0.571 (90%)	0.608 (90%)	0.604 (96%)	0.616 (95%)
SBS	1	**T7 HFD**	**T8 HFD**	**T7 HFD**	**T7 HFD**
	2	**AF4 HFD**	T8 PSDθ	**AF4 HFD**	**F7 svdEnt**
	3	**T8 HFD**	T7 PSDβ	F8 PSDγ	**F7 HFD**
	4	F7 PSDγ	T7 PSDγ	**T8 HFD**	FC5 PSDα
	5	FC5 PSDβ	P8 PSDβ	F8 PSDβ	FC5 PSDβ
		0.578 (90%)	0.598 (87%)	0.607 (94%)	0.603 (92%)

*__bold features__ were not used in the original AMIGOS set

It is interesting to note that the top feature in every case uses information from the temporal lobe. The temporal lobes have internal connections to the limbic system [53], whose primary functions includes emotion processing [53]. The temporal lobe, however, also plays a crucial role in processing the speed and frequency of sound, such as in speech, and in visual processing (such as facial recognition) [53]. Consequently, while the temporal lobe may be relevant for emotion processing, it may also be active due to the processing of the visual and auditory information from the movie.

3.6 LOMO-Within Cross-validation

Table 3.6 shows the results for the LOMO-Within cross-validation framework. This framework is the same as the LOMO-Inter case, but the models are trained with only a single user; the results are, therefore, participant-dependent. This framework evaluates how well a model trained for a particular user can extrapolate emotional information for a new movie (for that same user). Again participants 8, 24, and 28

were excluded as they did not participate in the long movie experiments, as were participants 17, 18, and 22 as they were missing annotations for some long movies. To reduce the impact of class imbalances for LOMO-Within, participants [20, 21, 23, 26, 31, and 38-40] were also excluded for valence and [11-13, 15, 20, 21, 23, 25-27, 29, 30, 35, 37, and 40] were excluded for arousal. Section 4.2 provides a more detailed analysis and discussion about class imbalance. A criterion of a minimum of ten high affect samples was set (empirically) for participants to be included in the LOMO-Within evaluation.

The LOMO-Within model was trained on one participant's data for three movies and tested on the same participant's data for the fourth movie. Results were averaged across 26 participants for valence and 19 participants for arousal. Features were selected based on the best average individual performances. As discussed in Section 2.4, participant-dependent models usually outperform participant-independent models because the inter-subject effect is mitigated and the model can learn patterns for a specific user.

Both a histogram and the Kolmogorov-Smirnov test were again used to test for normality, and since the data did not follow a normal distribution, statistical significance was again determined using the Kruskal-Wallis H-test. For both the SVM and LDA models, the feature selection techniques significantly improved the classification performance ($p < 0.01$). This reinforces the importance of feature selection to overcome the curse of dimensionality, particularly when selecting from many features with comparably little data.

As can be seen by comparing Tables 3.4 and 3.6, the LOMO-Inter results outperform the LOMO-Within results. This differs from the literature, where participant-dependent models typically outperform participant-independent models as they avoid inter-subject variability [160][42][74]. These models, however, are typically evaluated using k-Fold or LOO as cross-validation schemes [17][117][122], and not LOMO as

Table 3.6: LOMO-Within (Participant-Dependent) Results for Each Classifier (CLF) and Feature Selection (FS) Technique

CLF, FS	**Valence**		**Arousal**	
	F1	*#Features*	*F1*	*#Features*
LDA, SFS	**0.596***	20	**0.657***	20
LDA, SBS	0.592*	13	0.642*	20
LDA, ALL	0.445	217	0.470	217
SVM, SFS	0.593*	53	0.616*	35
SVM, SBS	0.578*	30	0.612*	26
SVM, ALL	0.481	217	0.528†	217

* denotes a significant difference from the CLF, ALL results ($p < 0.05$)
† denotes a significant difference between the LDA and SVM results ($p < 0.05$)

is used here. Using k-fold or LOO, some information about the stimulus (i.e., the movie) is used to train the classifier, making it easier to generalize the rest of the movie - sometimes even adjacent samples. In the LOMO scheme, this information is not present during training, reducing the observed performance, but making the evaluation results more realistic.

There is also a limited amount of data to draw from for training and testing for the LOMO-Within case. There are only four long movies in the AMIGOS dataset, so when one movie is left out, the training set is reduced further to three movies. When LOO or k-fold are used, the full movie may not be excluded from training, thus providing more exemplars to the classifier. These exemplars can be critical, especially when there are class imbalances such as in the AMIGOS dataset.

To reduce the impact of these imbalances, only participants with a minimum of ten high affect samples for each movie were used. Although ten samples still represent a small proportion of the overall data, to choose a higher number would have substantially reduced the number of testable models. This is a limitation of the AMIGOS dataset, and more films would be needed to obtain a better representation of LOMO-Within results.

From Table 3.6, it can also be seen that the number of features needed to reach the maximum performance is smaller than for other cross-validation models. This is likely because of the participant-dependence, eliminating the need to explain inter-subject variability. While SFS outperformed the SBS results for both SVM and LDA, the result was not statistically different for valence or arousal (p ranging from 0.617 to 0.889). As in LOMO-Inter, the SFS and SBS performance differences between SVM and LDA were also not statistically significant (p ranging from 0.564 to 0.773).

It can also be seen that the differences between the valence and arousal performances were greater than seen for the LOPO and LOMO-Inter frameworks. This is consistent with other works using the AMIGOS dataset, which tend to yield better arousal recognition than valence. This could indicate that EEG (or at least the features extracted here) holds more discriminative arousal information, or that the external annotators were better able to identify a person's state of arousal.

From Table 3.7, it is clear that HFD is again an important feature for the LOMO-Within framework. In all of the LDA models, SFS chose HFD as the top feature. Also, as with LOMO-Inter, the top feature in every case was derived from the temporal lobe, reaffirming its importance in emotion processing. In keeping with the smaller number of features selected for this case, groups of five features were able to achieve over 89% of the maximum F1-scores. The arousal SVM, SFS model required multiple iterations to find a meaningful feature set, resulting in a stagnant performance for the first several iterations. Iterations 13 to 17 increased the performance by 6%, and these iterations added T7 HFD, AF4 PSDγ, AF3 HFD, F7 HFD, and P8 HM. Not surprisingly, the HFD features and the temporal and frontal lobe channels greatly increased the performance. SVM using SBS does not choose any HFD features as part of the best-performing feature sets and instead chooses many more features from the AMIGOS [3] work. As HFD has been frequently chosen as a top feature, it likely improves the emotion recognition performance, thus its absence could have led to the

lower performance.

Table 3.7: Top Five LOMO-Within Features Selected by Feature Selection

		LDA		SVM	
		Valence	*Arousal*	*Valence*	*Arousal*
SFS	1	**T7 HFD**	**T7 HFD**	**T7 HFD**	**T7 PFD**
	2	AF4 PSDγ	FC6 PSDθ	T8 PSDγ	**P8 PFD**
	3	T8 PSDβ	AF3 PSD slowα	**O2 DFA**	**P7 PFD**
	4	**FC6 HFD**	AF3 PSDγ	**AF4 FI**	**F7 PFD**
	5	P Asymmγ	**T7 FI**	**AF4 PFD**	**O1 PFD**
		0.569 (95%)	0.606 (92%)	0.574 (97%)	0.549 (89%)
SBS	1	**T7 HFD**	**T7 HFD**	T8 PSDγ	T8 PSDγ
	2	AF3 PSDγ	**AF3 HFD**	**O2 HC**	T Asymmγ
	3	**T8 HFD**	AF3 PSD slowα	AF3 PSDγ	AF3 PSDθ
	4	T8 PSDθ	F4 PSDβ	T Asymmγ	P7 PSDγ
	5	**AF4 FI**	**F7 HFD**	F8 PSDγ	**P7 HC**
		0.560 (95%)	0.616 (96%)	0.539 (93%)	0.562 (92%)

*__bold features__ were not used in the original AMIGOS set

3.7 LOPMO Cross-validation

The LOPMO cross-validation evaluation framework gives an indication of how well a model can generalize to both a new person and a new stimulus. This makes it the most challenging, but also the most representative of potential real-world performance. The results using the LOPMO strategy are, therefore, reported in Table 3.8. The LOPMO model was created using the data for 34 participants (again participants 8, 24, and 28 were excluded, along with participants 17, 18, and 22 as with LOMO-Within). For each testing fold, the data from a single participant and movie, both unseen during training, were used to test the model. All data related to this movie were excluded from the training set (irrespective of participant) and all data related to this participant (irrespective of movie) were also excluded from training. Therefore, the training set consisted of the data for all other movies and participants. Features were

chosen as those that gave the maximum performance, averaged over the 136 results from each partition. Feature selection was not performed for the SVM classifier due to time constraints and the time-intensive nature of SVM.

Table 3.8: LOPMO (Participant- and Movie-Independent) Results for Each Classifier (CLF) and Feature Selection (FS) Technique

CLF, FS	**Valence**		**Arousal**	
	F1	*#Features*	*F1*	*#Features*
LDA, SFS	0.575	61	0.670*	102
LDA, SBS	**0.594***	28	**0.686***	72
LDA, ALL	0.521	217	0.605	217
SVM, ALL	0.511	217	0.574	217

* denotes a significant difference from the LDA, ALL results ($p < 0.05$)

Of all the frameworks, LOPMO had the greatest difference between the performances of valence and arousal. Arousal was recognized with similar success as it was when using the LOMO-Inter and LOPO strategies. Valence, however, saw a decrease in performance. Although many factors could contribute to this, it could suggest that the external annotators had a harder time understanding valence based on visual facial expression. After failing the Kolmogorov-Smirnov test and the visual histogram test for normality, a Kruskal-Wallis H-test was used to determine significance. Again, the benefit of feature selection was supported by the statistically significant differences between using all 217 features and the chosen subset ($p < 0.02$). No statistical significance was found, however, for all 217 features and the SFS features using LDA for valence ($p = 0.094$).

Table 3.9 shows the top five features selected for SFS and SBS. When classifying valence using SFS, the top five features obtained only 84% of the maximum F1-score for that condition. This is in contrast to previous results from the other evaluation frameworks, which described over 90% of the maximum F1-scores.

Table 3.9: Top Five LOPMO Features Selected by Feature Selection (LDA)

	SFS		SBS	
	Valence	*Arousal*	*Valence*	*Arousal*
1	**FC6 HM**	**T8 HFD**	**AF4 HFD**	**T8 HFD**
2	F8 PSDα	T8 PSD slowα	**T7 HFD**	T8 PSD slowα
3	**O1 DFA**	**AF3 svdEnt**	**T8 HFD**	T7 PSDβ
4	AF Asymmθ	**T7 HFD**	AF3 PSDγ	T7 PSDγ
5	P Asymmθ	**F4 HFD**	P8 PSDγ	AF3 PSDα
	0.484 (84%)	0.630 (94%)	0.544 (92%)	0.626 (92%)

*__bold features__ were not used in the original AMIGOS set

The features chosen within the LOPMO framework also differ from those of LOMO-Inter and LOMO-Within, as more features from lobes other than the temporal lobe are introduced sooner. When classifying with LDA, features from the frontal lobes were chosen first. However, like the temporal lobe, the left frontal lobe plays a crucial role in language comprehension [53]. The frontal lobe is also important in emotion retrieval, specifically as actions unfold over time [70].

It has been mentioned that SBS may be better able to understand the relationships between features because it begins with groups of features. Figure 3.4 demonstrates how SBS is better able to retain meaningful groups of features for fewer features, whereas SFS requires multiple iterations before finding groupings that lead to improvements.

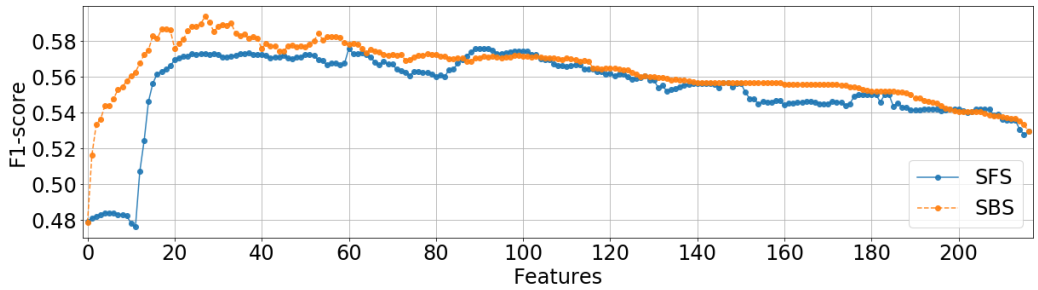


Figure 3.4: Comparison of SFS and SBS for Valence, LOPMO Validation Using LDA Classification

3.8 Summary of Results

The results from the various cross-validation metrics showed the benefit of feature selection and that SFS often does not perform as well as SBS. This is likely because there are important relationships between EEG-based features that are only captured by SBS. For both feature selection techniques, the first five features provided around 90% of the maximum information to the classifier. Also, the additional features extracted in this work were frequently chosen by the various models, improving upon previously reported results. Of the extracted features, HFD was chosen most frequently as the first feature. Alarcão's review of emotion recognition using EEG signals [18] found that the most commonly used features in the literature include PSD (22.2%), and entropy (15.9%), whereas fractal dimensions are only used in 7.9% of works. As documented in [149], although the Fast Fourier transformation (FFT), used for PSD, is beneficial for the analysis of stationary signals, neurophysiological processes are often non-stationary. This work corroborates those of Anh et al. [159] and Solhjoo et al. [83] that HFD is informative for EEG emotion recognition tasks.

The chosen features also align with reports from neuropsychology, with temporal lobe channels frequently being chosen as a top feature, followed by frontal lobe channels. The temporal lobe is critical for this type of emotion recognition as it plays a role in audio, visual and emotion processing. The frontal lobes play key roles in emotion retrieval, as well as audio processing; therefore, it makes sense that the majority of the chosen channels come from these lobes.

The best results from the various cross-validation techniques are tabulated in Table 3.10. Of all the cross-validation techniques, LOPO performed the best. LOMO-Within performed the worst for arousal and LOPMO performed the worst for valence, yet both still outperformed the results from AMIGOS, likely due to the additional features extracted as part of this work. These results again show how arousal usually

outperforms valence, corroborating that valence is more difficult to quantify [161].

Table 3.10: Best Results from Cross-validation Techniques

	Valence			**Arousal**		
	F1	*CLF*	*FS*	*F1*	*CLF*	*FS*
AMIGOS [3]	0.557	NB	FLD	0.571	NB	FLD
LOPO	0.706	SVM	SBS	0.699	LDA	SBS
LOMO-Inter	0.648	SVM	SBS	0.684	LDA	SBS
LOMO-Within	0.596	LDA	SFS	0.657	LDA	SFS
LOPMO	0.594	LDA	SBS	0.686	LDA	SBS

The best performing results that are shown in Table 3.10 come from both SVM and LDA classifiers. Though both were implemented for each feature selection technique and affect type, there were no statistically different performances once feature selection had been completed. The creation and testing of SVM models, however, takes more than 10 times that of LDA, (e.g., almost two weeks to compute the valence, LOPO, SBS results, on an Intel® Core™ i9-7920X CPU @ 2.90GHz with 24 concurrent workers). For this reason, LDA was adopted for the remainder of the work.

While the participant-dependent analysis using LOMO-Within did not outperform the independent model of LOMO-Inter, the reasons this could have occurred include the chosen cross-validation technique, the small amount of data, or the imbalance in the labels. Chapter 4 extends upon the results of this chapter by considering the fidelity of the labels created as part of AMIGOS.

Chapter 4

Training Strategies

To better understand the results obtained in Chapter 3, this chapter presents a series of different explorations. To begin, the impact of window length and increment are briefly investigated. Because classification results are so heavily dependent on the ground truth labels against which they are evaluated, the AMIGOS labels are evaluated. The impact of gender is studied in response to gender differences and biases in the dataset. Finally, a potential framework, in the form of rejection, is proposed to improve the performance of the system in a real-world application context. The results presented in this chapter reflect those of an LDA classifier for the reasons outlined in Section 3.8.

4.1 Window Size & Increment

Despite the use of time-evolving labels in AMIGOS, the use of non-overlapping twenty-second labels limits its utility in understanding temporal dynamics. Although these windows are assigned a single label, because they are relatively long in comparison to possibly rapid emotional responses to audio-visual stimuli (e.g., laughing at a joke, or jumping in response to a surprise), they may contain multiple emotions or levels of emotion. Also, the creation of a decision every twenty seconds may lead to large observed swings in affect from window to window. Effectively, the signal (emotion) is under-sampled by this labelling approach. Had there been greater label and decision granularity, it is possible that more information could have been learned about how emotion evolves throughout the movies.

To explore this concept, additional labels were created in increments of one second, using linear interpolation of the original annotations.

$$L'_{i,N} = L_i + \left(\frac{N}{20}(L_{i+1} - L_i)\right) \qquad for\ N \in \{1, 2, \ldots, 19\}\ \&\ i \in \{0, 1, \ldots, n-1\}, \tag{4.1}$$

where $L_0 = 0$ and $L_1, \ldots, L_n$ are the AMIGOS labels starting at the 0-20 second label to the second to last label (last full non-overlapping label). The term N represents the time (in seconds) after the label L_i and before L_{i+1}.

An example of newly created valence labels is shown in Figure 4.1 for one participant watching one movie, with the original labels denoted by the larger (red) markers.

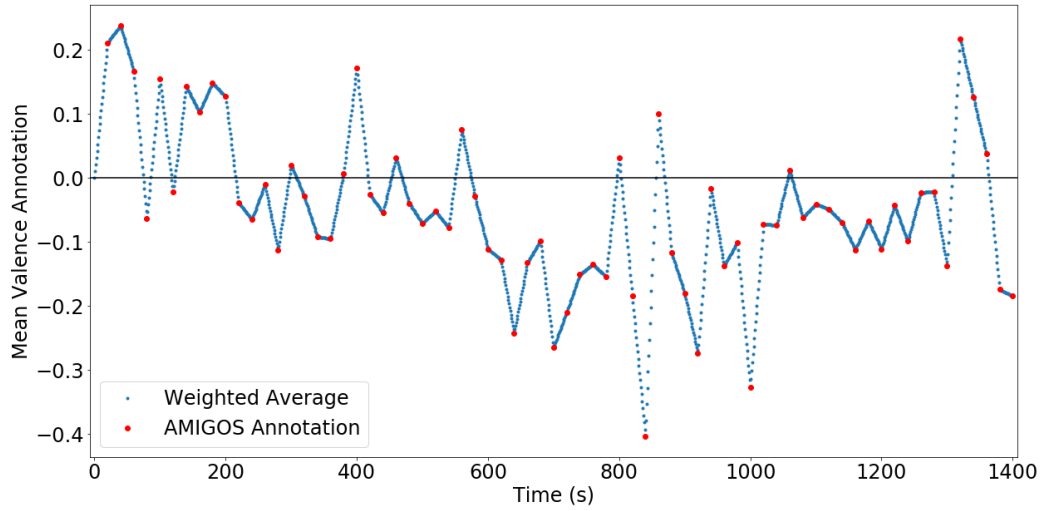


Figure 4.1: Linearly Interpolated Valence Labels for a Participant Watching *The Descent*

Using these new labels, the impact of window increment and length was explored. To facilitate comparisons between different lengths, windows were aligned as shown in Figure 4.2. That is, decisions corresponding to a particular label used the preceding N seconds of data, where N was dictated by the window length. Correspondingly, no decisions were made for the first 59 seconds of any movie (because the 60-second

windows were not made until the 60th second). Decision windows were incremented by the same amount from there on, ensuring alignment for all decisions.

Windows ranging from 10 to 60 seconds were explored based on pilot results; however, windows that are too long overlap multiple labels, and windows that are too short may not include sufficient data to make an appropriate decision about emotion [21].

Figure 4.3 shows the effect of window length on the computed F1-scores for the LOMO-Inter case. Comparisons show the performance of SBS and SFS for one and twenty-second window increments for various window lengths, for both arousal (above) and valence (below). It should be noted that the values in the legend refer to the increment sizes, whereas the horizontal axis denotes the window length. In general, as was found in Chapter 3, SBS tended to outperform SFS, and arousal performed better than valence. While there may be a small trend towards a lower performance at higher and lower window lengths, these results suggest that the twenty-second window size, corresponding with the AMIGOS label intervals, performed best in this evaluation framework. Caution should be taken when interpreting these results, however, as they likely reflect the level of agreement with the method of labelling, and not necessarily the information content of the EEG signal itself.

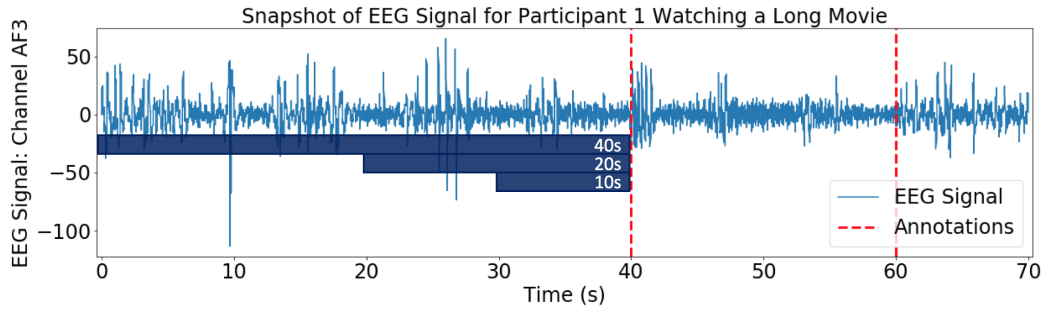
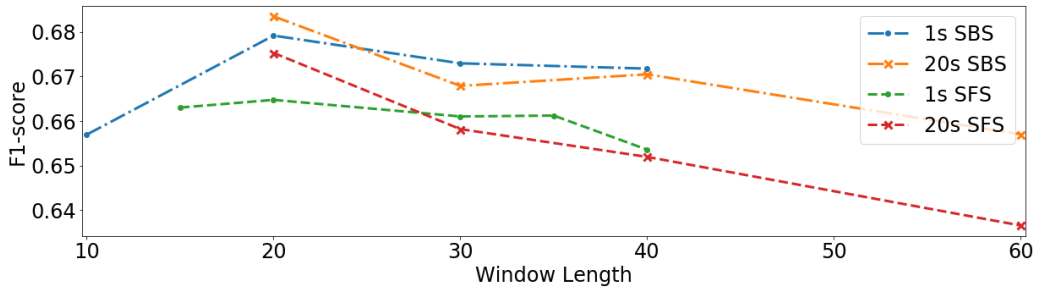


Figure 4.2: Alignment of Windows for Classification

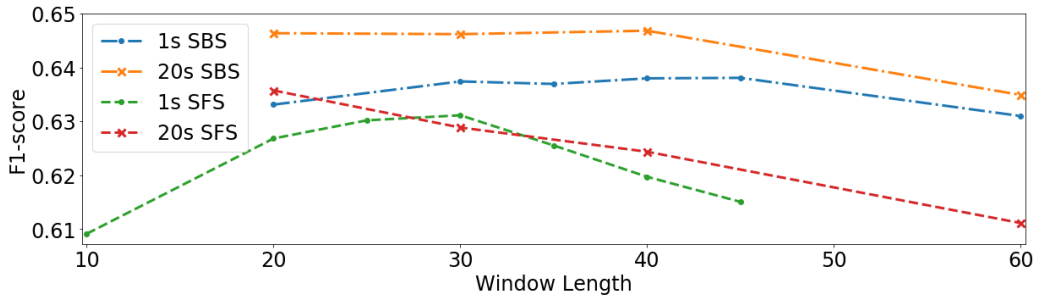
The weighted averaging interpolation method also did not provide much benefit, with the twenty-second increments outperforming the one-second increments. From

inspection of Figure 4.1, it can be seen that a higher-order polynomial fit could better model the temporal dynamics; however, the poor resolution of the original labels may not support this type of approach in general. Future work would benefit from re-labelling of the video data with a higher temporal resolution; however, this substantial task remains outside of the scope of this work.

Given the results of this exploration, the remainder of this work continues to use a twenty-second increment and a twenty-second window.



(a) Arousal



(b) Valence

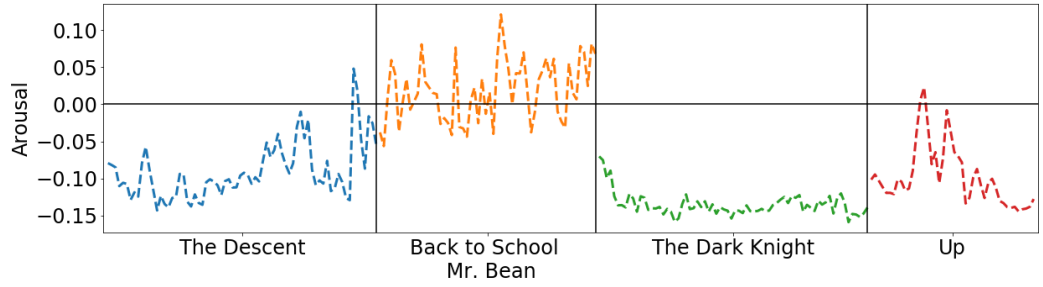
Figure 4.3: LOMO-Inter Performance vs Window Length For Different Window Increments When Using Linearly Interpolated Labels

4.2 Labels

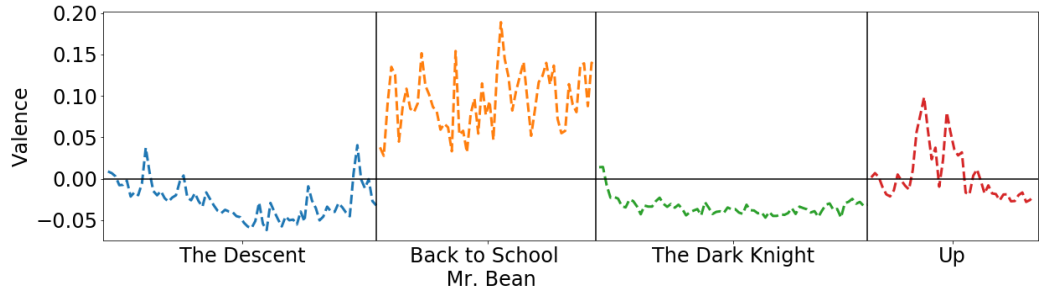
Investigation of the distribution of AMIGOS labels reveals a bias towards low affect classes, causing an imbalance in the class labels. Class imbalance in machine learning can substantially affect the training, and consequently, the performance of

classification models [162, 125, 163].

Figure 4.4 shows the mean of the external annotations across all participants as they watch the long films. As can be seen, many of the samples are below zero and are thus categorized as low affect for the classification task. Worse, the distribution of high affect and valence is heavily non-uniform across movies. Of the four movies, *Mr. Bean* is the primary source of high arousal and valence information, making it problematic in a leave-one-out cross-fold validation evaluation framework. For instance, when that movie is left out as the testing movie, there is very little high arousal/valence information from which to inform the classifier training. This severe imbalance across the movies could impact any of the evaluation frameworks, and particularly those that evaluate generalization to new movies (LOMO-Within, LOMO-Inter and LOPMO). It also means that the true variance of high affect emotion is likely not captured in this dataset, as almost all examples come in response to one specific movie.



(a) Arousal



(b) Valence

Figure 4.4: Mean External Annotations for the Long Movies in AMIGOS

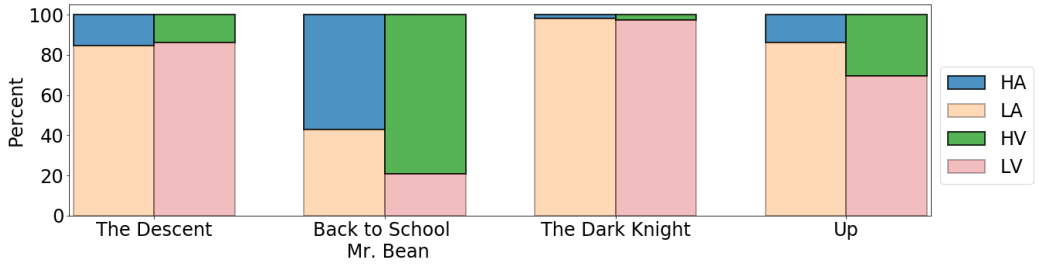
The number of samples from each class was computed and the distribution of classes is shown in Table 4.1. It can be seen that, although the originators of the AMIGOS dataset chose these films to evoke emotions in the four quadrants of emotion, the labels show that there is a strong bias towards the low-valence low-arousal (LV-LA) quadrant (with 68% of the samples).

Table 4.1: Distribution of Arousal and Valence Labels, Averaged Across All Annotators and Long Movies

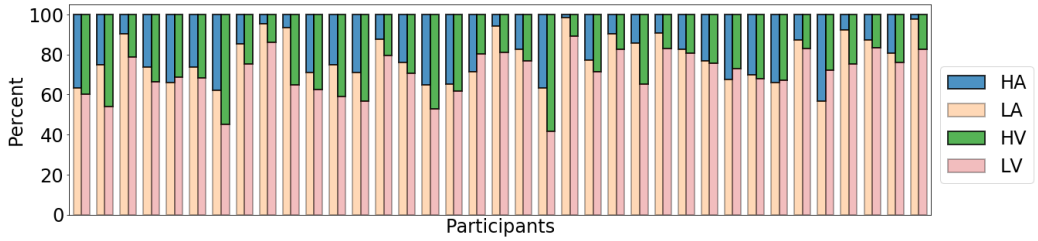
	LV	**HV**
LA	68%	11%
HA	3%	18%

Additionally, breaking these labels into the two sets of two-class problems (high vs. low valence and high vs. low arousal), shows that low valence represents 71% of all valence examples and that low arousal represents 79% of all arousal examples.

Further evaluating this imbalance by movie, as shown in Figure 4.5a, shows that *Mr. Bean* has the most high affect samples, with 63% high valence samples and 64% high arousal samples. *The Dark Knight* has the most low affect samples, with 97% low valence samples and 98% low arousal samples. The participant-specific breakdown of these results is shown in Figure 4.5b. It can be seen that there are eight participants with less than 10% high arousal, with one participant (#23) being labelled as high arousal for only 1% of all arousal samples. Valence is slightly more balanced than arousal but still tends towards low valence samples. The lowest percentage of high valence is 11% (again for participant 23). For the cross-validation schemes where a movie is left out, especially for LOMO-Within, it is possible for no high affect samples to be included in the training set if this imbalance is not considered. As in Chapter 3, a minimum of ten high affect samples was, therefore, chosen for the LOMO-Within training set to ensure that at least some high affect data were seen during training. This does not address the imbalance issue but was necessary for training purposes.



(a) Average Breakdown of Valence and Arousal Samples by Movie



(b) Breakdown of Valence and Arousal Samples by Participant

Figure 4.5: Class Breakdown on a Per Movie and Per Participant Basis

Although there is a clear overall bias towards low valence and arousal, these results represent only the mean values across three external annotators. To better understand how these biases were achieved, and the degree of alignment between annotators (and thus, confidence in their labels), the external annotations of the individual annotators were investigated.

Three different annotators watched each twenty-second clip (in randomized order) and provided valence and arousal labels, accordingly. Figure 4.6 shows the range for each annotator's annotations for valence and arousal. It can be seen that, while the values were allowed to range from -1 to 1, the majority of the annotations were between -0.5 and 0.5. This could indicate a degree of uncertainty in the labels, suggesting that annotators were unable to confidently infer emotion from the videos of participants' faces. It can also be seen that Annotator 1 tended heavily towards negative Arousal, whereas Annotator 2 tended towards positive Arousal. Given that these labels were obtained from the same dataset, this casts doubt on the fidelity of the labelling process.

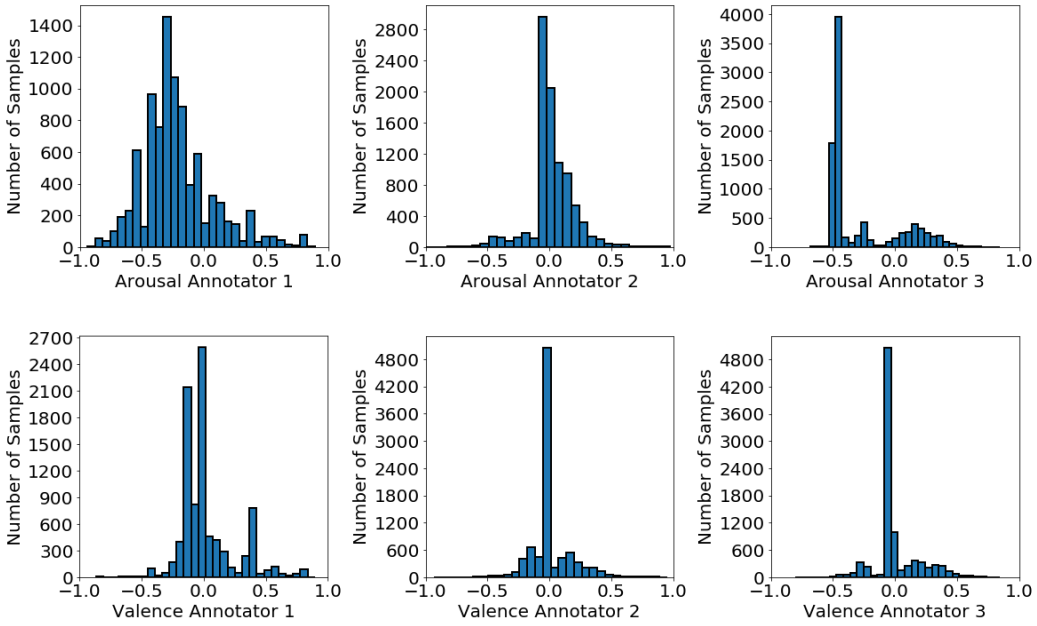


Figure 4.6: Distribution of the Individual Valence and Arousal Annotations

To better understand the impact of inter-annotator disagreement, the classification performance of the system was evaluated when using each individual annotator's labels as the ground truth. The feature sets determined in Chapter 3 using either SFS or SBS were used to train the LDA classifier, and new models were trained and evaluated using each annotator's labels. Table 4.2 shows the results for the valence and arousal problems within the LOMO-Inter cross-validation framework.

Table 4.2: LOMO-Inter Results When Using Annotations from Individual Annotators and Their Mean

Affect	Scheme	A1	A2	A3	Mean
Valence	LDA, SFS	0.544	0.563	0.590	0.636
	LDA, SBS	0.560	0.577	0.599	0.646
Arousal	LDA, SFS	0.543	0.550	0.592	0.675
	LDA, SBS	0.539	0.551	0.590	0.684

As can be seen, performance decreases when the labels for each individual annotator are used instead of their mean. There is also some difference in the performance

per annotator, with annotator 1 being the worst and annotator 3 being the best. The distribution of classes across annotators is also shown in Table 4.3, confirming consistent but differing degrees of bias across annotators.

Table 4.3: Breakdown of Class Imbalance Per External Annotator (1/2/3)

	LV	**HV**
LA	65/57/71%	18/1/5%
HA	4/16/4%	13/26/20%

Despite the differences in individual annotator performances, the fact that their mean yielded the best performance suggests that averaging was able to smooth out disagreement and reduce the impact of outlier labels. It is, therefore, possible that recruiting additional annotators to evaluate the dataset could further improve the accuracy of these 'ground truth' labels.

The impact of annotator agreement was further explored by evaluating classification performance only on samples for which there was class agreement between the three annotators. Models were retrained and tested using only these samples. Table 4.4 shows these results for the LOMO-Inter validation framework for valence and arousal (shown as *Agree*). To understand if this restriction changes the chosen features, feature selection was recomputed (shown as *Agree New*). For comparison, the Chapter 3 results using all of the samples (shown as *All Samples*) is also listed. The impact of the *All Samples* model on only agreeing test data was investigated as well (shown as *Test Agree*).

Table 4.4: LOMO-Inter Classification Results When External Annotators Agree on the Sample's Class

	Valence				**Arousal**			
	Agree	*Agree New*	*All Samples*	*Test Agree*	*Agree*	*Agree New*	*All Samples*	*Test Agree*
SFS	0.605	0.623	0.636	0.576	0.637	0.682	0.675	0.600
SBS	0.625	**0.649**	0.646	0.591	0.645	0.683	**0.684***	0.597

* denotes a significant difference between the *All Samples* and *Test Agree* results ($p < 0.05$)

When considering only the assigned binary class labels, the annotators were found to agree 75% of the time for valence and 61% of the time for arousal. This is quite poor, considering this is used as the ground truth for EEG classification. Furthermore, the majority of samples that were agreed upon fell into the low affect classes, leaving heavy disagreement in the high affect cases. It should be noted that in removing labels that weren't in agreement, the valence and arousal imbalances further increased by 7% and 6%, respectively (in absolute terms). Figure 4.7 provides a visual representation of the agreeing samples. The reduced number of overall samples can be seen (the bars don't stack to 100%) along with the reduced proportion of high affect samples.

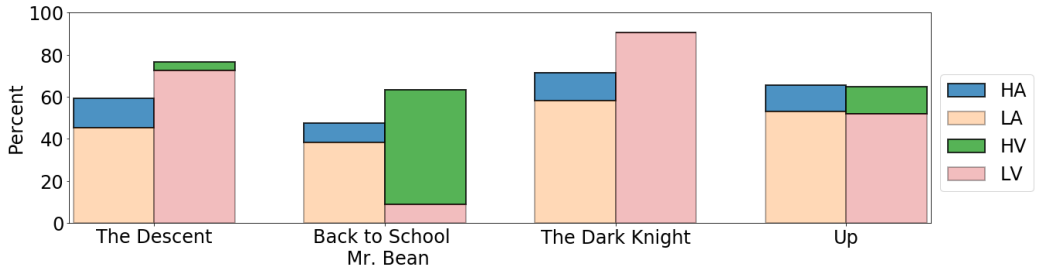


Figure 4.7: Agreeing Valence and Arousal Samples by Movie

The results of Table 4.4 show that performance decreased when using the agreed-upon subset of labels. When feature selection was recalculated, results were comparable (a Kruskal-Wallis test determined one significant difference, $p < 0.05$, between

All Samples and *Test Agree* for the Arousal, SBS model). While it was expected that the performance would improve when using only agreed upon annotations, this result was outperformed by *All Samples*. Importantly, however, using only these samples reduced the amount of available data and increased class imbalance, yielding fewer high affect samples with which to train the model. To better understand the role of annotator agreement, a larger and more balanced training set would be needed to ensure that both high and low affect samples are equally represented when the model is created.

4.3 Training Techniques

With a better understanding of the limitations of the AMIGOS dataset and labels, a variety of different training techniques were explored to improve the classification performance. These included examination of gender biases, modifications of the labels to remove bias, and classification based on decision confidence, as outlined as follows.

4.3.1 Gender Bias

In addition to the affect bias seen previously, the 40 participants in the AMIGOS dataset are composed of 13 females and 27 males, resulting in a gender-biased dataset. Studies have suggested the possibility of gender-related neural differences in response to emotional stimuli [164, 165, 166], and so the dataset was disaggregated based on gender. Separate female and male models were trained and tested using the results of the SBS feature selection for all four cross-validation schemes outlined in Chapter 3. A Kruskal-Wallis H-test was used to compare the differences between the disaggregated and mixed models, with a p-value of less than 0.05 considered significant.

Table 4.5 shows the results of this gender analysis. Overall, the female models outperformed the male models by approximately 6% and 3% for valence and arousal,

respectively. Many differences showed strong, if not significant, trends. As compared to the mixed-model case, the female model had better performance throughout (up to 7%), while the male model was always within 1% of the mixed model results. Although only the female arousal LOMO-Within model showed a significant difference over the male model ($p = 0.047$), other models trended heavily towards significance, such as the female valence LOPMO model ($p = 0.052$ vs the mixed model, and $p = 0.065$ vs the male model).

Table 4.5: Results from Gender Disaggregation (LDA SBS)

CV	**Valence**			**Arousal**		
	Female	*Male*	*Mixed*	*Female*	*Male*	*Mixed*
LOPO	**0.769**	0.701	0.695	**0.731**	0.702	0.699
LOMO-Inter	**0.707**	0.639	0.646	**0.716**	0.689	0.684
LOMO-Within	**0.630**	0.600	0.592	**0.700***	0.614	0.642
LOPMO	**0.657**	0.593	0.594	**0.712**	0.693	0.686

* denotes a significant difference from the corresponding Male results ($p < 0.05$)

To evaluate whether some of the observed improvement could be due to differences in class balance, the gender-disaggregated class breakdown was evaluated. High arousal was found to represent 27% of the data for females and only 18% of the data for males. Similarly, high valence represented 32% for females and 27% for males. Compared to the mixed-gender model (21% and 29% for high arousal and valence, respectively), it can be seen that the class breakdown became more imbalanced for men, and more balanced for women. It is possible that this improved balance contributed to the better female results, but it should be noted that the classes remained quite severely imbalanced.

A new SBS feature selection was also conducted for each gender model to determine whether there may be features that work better for male and/or female participants. The performance using the best female, male, and mixed features were

then compared for each model, as shown in Table 4.6. The F1-scores are shown for three cases: *Same*, the best set of features determined using the gender being tested, *Other*, the best set of features determined when using the other gender (e.g., females being tested using the best male features), and *Mixed*, using the best set of features as determined using the mixed model.

Table 4.6: Effect of Feature Selection for Gender Disaggregation (LDA SBS)

CV	**Model**	**Valence**			**Arousal**		
		Same	*Other*	*Mixed*	*Same*	*Other*	*Mixed*
LOPO	F	**0.769**	0.645	0.665	**0.731**	0.653	0.653
	M	0.701*	0.607	0.669	0.702*	0.624	0.681
LOMO Inter	F	**0.707***	0.618	0.647	**0.716**	0.633	0.677
	M	0.639	0.576	0.605	0.689*	0.611	0.653
LOMO Within	F	**0.630***	0.547	0.612	**0.700***	0.550	0.685
	M	0.600*	0.489	0.578	0.614*	0.485	0.595
LOPMO	F	**0.657***†	0.508	0.533	**0.712***†	0.590	0.602
	M	0.593*	0.505	0.569	0.693*	0.605	0.657

* denotes a significant difference between the *Same* and *Other* results ($p < 0.05$)
† denotes a significant difference between the *Same* and *Mixed* results ($p < 0.05$)

As an example of the differences in the chosen features, Table 4.7 highlights the top five features selected for the LOPO valence case for each model. As in Chapter 3, the temporal lobe and HFD remain important for all the models, but the chosen features vary considerably. Importantly, many of the features chosen in all cases were previously unreported with the AMIGOS dataset.

Table 4.7: Top Five LOPO Valence Features Selected by SBS for Gender

	Female	*Male*	*Mixed*
1	T7 PSDγ	**T7 HFD**	**T8 HFD**
2	**F7 HFD**	**AF3 svdEnt**	**AF4 HFD**
3	**AF4 HFD**	AF3 PSD slowα	**F7 PFD**
4	FC5 PSDβ	P7 PSDα	**FC5 HM**
5	**F3 SpecEnt**	**T7 FI**	**O2 HC**

*__bold features__ are new features not previously reported with the original AMIGOS works

These results support previous findings in the literature that emotion information in the EEG signal may be, at least partially, gender-specific. It can be observed that using the features chosen for another gender significantly decreases performance as compared to those selected for that gender. A Kruskal-Wallis H-test found significant differences as compared to the other model in almost all cases. With additional data, some other models have yielded significant differences as well (e.g., female valence LOPO yielded $p = 0.057$ *vs.* the male model and $p = 0.065$ *vs.* the mixed model). The improvements when using a gender-specific model tend to be larger for females, consistent with suggestions that women may share more similar EEG and or facial expression patterns, thus reducing their inter-subject variance [165]. Conversely, it is also possible that the presence of the male-bias in the dataset (and thus the mixed case) contributes to the smaller difference from the male-specific case. Again, the impact of the imbalanced amount of data in each case, and the reduced overall amount of data, may also factor into performance differences.

4.3.2 Classification Thresholds

Given the difficulty of emotion recognition, it is common to simplify the problem to one of discrete affect classification, as seen through this dissertation. To map from the continuous valence and arousal values of the labels, a zero threshold is used to

assign a label into the high or low affect class. As discussed, however, this produces a biased dataset for AMIGOS. As was seen in Figure 4.4, many of the labels fall below this zero threshold.

To determine if there may have been a persistent bias in the labels that was impacting the classification performance, two strategies were explored. The first approach consisted of shifting the previously zero threshold to the median of the classes to create a 50/50 balance. For valence, this amounted to a threshold shift to -0.02 (2% of the range) and for arousal, to -0.12 (14% of the range). Secondly, because both valence and arousal suffered from a bias, as shown in Figure 4.8, a zero-mean normalization was attempted. In this method, the thresholds were shifted to the mean of each of the arousal and valence label distributions. For valence, the threshold was correspondingly set to +0.005 (1% of the range) and at -0.080 (10% of the range) for arousal, shifting the imbalance from 29% high valence and 21% high arousal to 27% high valence and 33% high arousal.

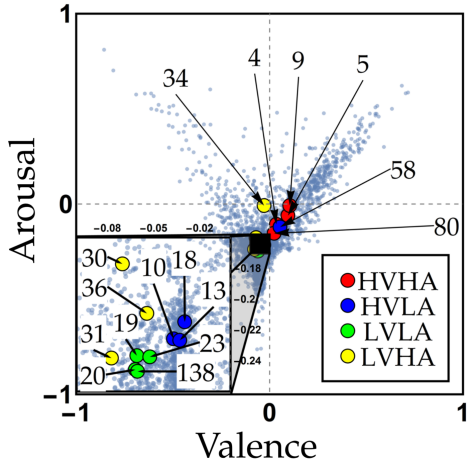


Figure 4.8: Distribution of External Annotations in the AMIGOS Dataset [3]

After changing the class designation thresholds, the data were re-classed using the original features found in Chapter 3, and classification training and testing were conducted as before. The LOMO-Inter cross-validation framework was used to ex-

amine the impact of the change. If the imbalance in the classes was due to a true, and erroneous, bias in the labelling (e.g., due to annotator inability to accurately infer emotion from the video), the results of this cross-validation would likely have yielded improved results. As shown in Table 4.8, which shows the results for each case, however, the balanced labels produced worse performance than the unadjusted labels. With the mean-adjusted approach, the only improvement came in the valence case. It should be noted, however, that this case introduced additional imbalance in the classes, likely confounding any real improvement. Statistical significance was determined using a Kruskal-Wallis H-test ($p < 0.05$).

Table 4.8: Results from Shifting Threshold (LOMO-Inter)

Threshold	**Scheme**	**Valence**	**Arousal**
Original	LDA, SFS	0.636	0.675*
	LDA, SBS	0.646	**0.684***
Balanced	LDA, SFS	0.529	0.584
	LDA, SBS	0.547	0.598
Mean	LDA, SFS	0.633	0.619
	LDA, SBS	**0.647**	0.633

* denotes a significant difference from the *Original* and the corresponding *Balanced* results ($p < 0.05$)

When the threshold was shifted, classification performance decreased. The only case of marginal improvement occurred when the imbalance increased. As the majority of the samples were labelled as low affect, it could be that the classifier was able to detect low affect well, and thus detect some incorrectly labelled high affect samples as low affect. When the thresholds were made smaller, however, low affect labels were then represented as high affect. The small number of high affect labels could mean that the classifier was unable to create a strong model for these samples, particularly for those around the threshold. To better determine the impact of shifting thresholds, more true high affect samples would likely be needed.

4.4 Selective Classification

This chapter has highlighted some limitations of the AMIGOS dataset, including a male bias, severe class imbalance, and questionable confidence in the labels used as ground truth for algorithm development and evaluation.

Importantly, the target labels were created by external annotators based on facial expressions. Facial expressions have been reported to be a poor indicator of what individuals feel both across cultures [167], and within cultures [168]. As was stated by Barrett et al. [169], "the validity of the conclusions that scientists draw about emotions depends on the validity of their initial assumptions." The current system is, therefore, trained to understand the consensus of what the AMIGOS external annotators believe the participants to be feeling, and not necessarily what the participants are actually feeling.

The lack of high affect labels has been thoroughly discussed in this chapter, and this imbalance could come from a few factors including the small number of long movies, the genres chosen for the long movies, and the selection and length of the clips from them. An imbalance in the training label impacts classification performance, as high affect might not be fully understood due to its inadequate representation in training data. The imbalance could also be due to the rating scale used by the external annotators. Several known challenges exist for rating scales, including using the scales inconsistently [170, 171], and the usage of only a small subset of the available range [172]. From Figure 4.6, it can be observed how little of the full range is used.

To understand how annotator confidence may impact model accuracy, the model performance distributions were compared back against the continuous label values. Correspondingly, Figure 4.9 shows the distributions for the correct and incorrect predictions for valence and arousal for the LOMO-Inter SBS model. The baseline features found in Chapter 3 were used. From Figure 4.9, it can be observed that

incorrect predictions tend to cluster more closely around the zero threshold than do the correct predictions.

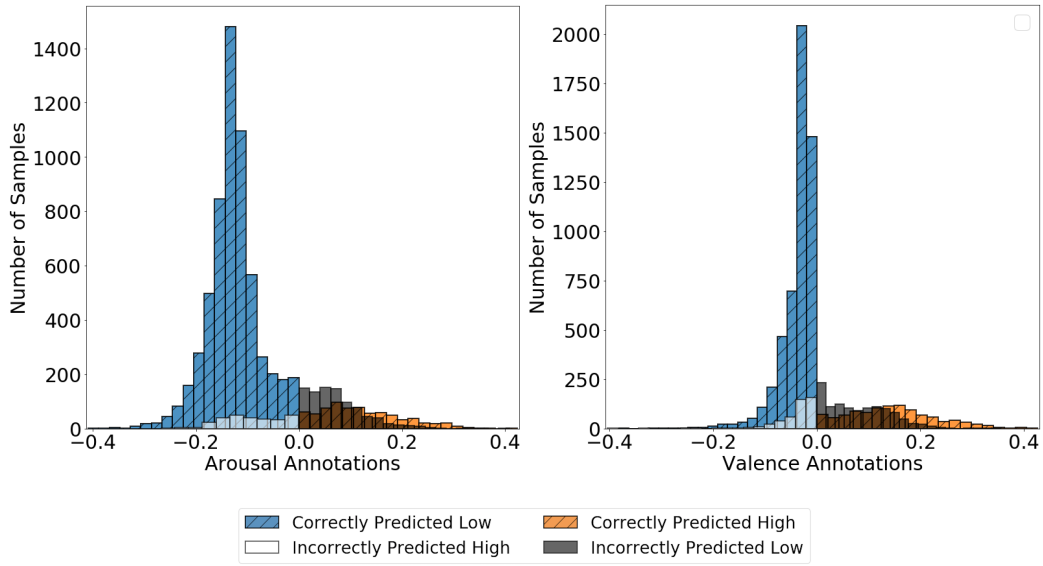


Figure 4.9: Distribution of the Correctly and Incorrectly Classified Valence and Arousal Annotations (LOMO-Inter SBS)

To further explore the uncertainty around zero, the F1-scores for various bins of continuous label values were calculated. The bins separated the continuous annotations from 0 to 0.2 in increments of 0.05 and then from 0.2 to 0.4 due to the disproportionate distributions of decisions in this bin. To facilitate consistent calculation of the F1-score, the ranges indicate the absolute values, meaning that a bin includes data of similar magnitude from both the high and low affect classes (e.g., from -0.4 to -0.2 and from 0.2 to 0.4). Figure 4.10 shows the F1-scores for the various bins of valence and arousal using the LOMO-Inter SBS model. Four different percentage values are reported within each bin in the figure. The top percentages show the classifier's average confidence for each class (high/low affect). Confidence values were determined as the average probability (*predict_proba*) output by the sklearn implementation of the LDA classifier for each corresponding decision. The confidences were mapped to each sample's class, and the average of the confidences for each class

in each bin are reported. The bottom percentage shows the proportion of the total samples within each bin (high/low affect).

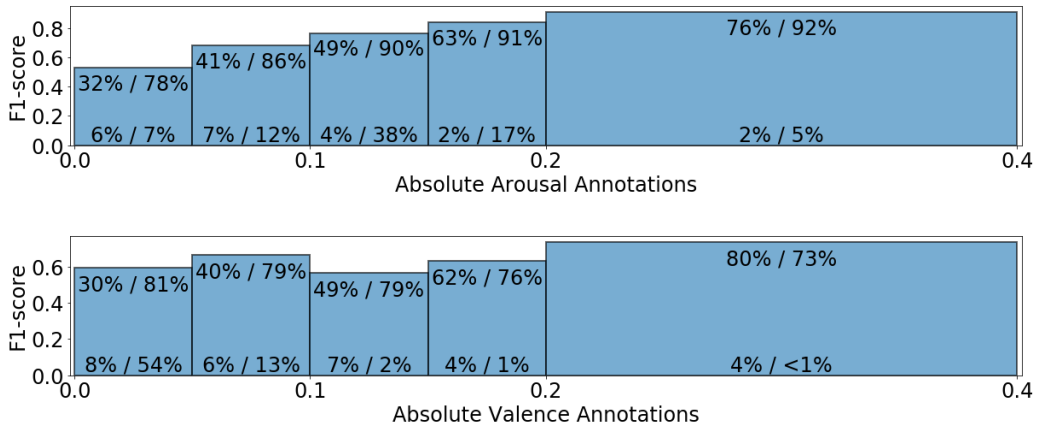


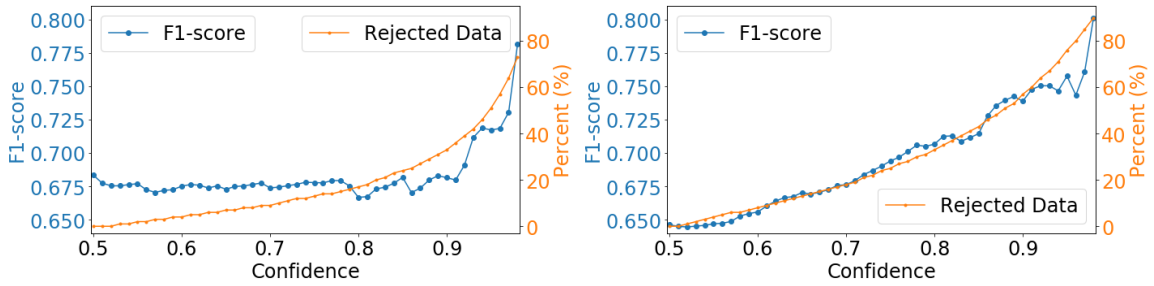
Figure 4.10: F1 Performance For the Mean External Valence and Arousal Annotations in Bins (LOMO-Inter SBS). The top percentages denote high/low affect confidence. The bottom percentages denote the proportion of high/low affect samples.

This figure demonstrates some of the insights that have been realized throughout this chapter. Confirming Figure 4.6, the highest magnitude continuous annotations bins have the smallest amounts of data. The imbalance in the dataset can also be seen (particularly for the 0.1-0.15 bin of arousal and 0-0.05 bin of valence), as was discovered in Section 4.2. As in Figure 4.10, it can be seen that there is a decrease in performance closer to the zero threshold. This suggests either that the classifier is unable to correctly differentiate within this region or that there could be errors in the labels. The classifier's confidence can be observed to generally be higher for the low affect data, likely due to the bias in the dataset. Near the zero threshold, the high affect confidences are particularly low, possibly suggesting that they are conditioned by uncertainty in the annotators' assessments of high affect.

The previous results largely compared classification performance with the continuous annotator labels. However, upon inspection of Figure 4.10, it can also be seen that F1-score tends to be higher for bins with higher classifier confidence. This behaviour could lend itself to a *rejection*-style framework, wherein only decisions with

high confidence are accepted. Such a framework would potentially reduce the number of samples wherein a decision is made, but may improve the performance of those decisions. Decisions were, therefore, *rejected* if a classifier's confidence was lower than a given threshold.

Figure 4.11 shows the resulting F1-score (left) as a function of confidence threshold. The percentage of samples that are rejected at those thresholds (right) is also shown. It can be seen that valence performance increases somewhat linearly with the rejection threshold, improving from 0.65 to 0.73 when 50% of samples are rejected. Although arousal can be seen to respond much more slowly, it too increases from 0.68 to 0.72 when 50% of samples are rejected. Interestingly, the (im)balance of classes appeared to remain relatively consistent between the rejected and non-rejected decisions as the threshold was varied.



(a) Arousal

(b) Valence

Figure 4.11: F1-Score vs Classifier Confidence (Left), and % of Samples 'Rejected' with Confidence Below a Set Confidence Threshold (Right) for the LOMO-Inter, SBS Case.

It can be seen that F1-score (in blue) increases with decision confidence, although more linearly for valence than for arousal.

Although this rejection framework incurs a substantial reduction in decisions made, it is possible that a given application may not need to make continuous decisions. For example, a mental health monitoring system may only need to track large changes or swings in emotion between high and low affect. Such an approach may

also facilitate more fine-tuning of error tolerance in affective computing systems. Because the confidence computed here used the default output from the LDA classifier implementation used in this work, it is possible that more elaborate approaches may yield a more favourable tradeoff between performance gain and rejection frequency.

Chapter 5

Discussion & Conclusion

Mental health challenges are a growing problem globally but have been shown to respond well to early intervention. Detecting consistent or large swings in emotion could be an important indicator as part of a prevention program. Affective computing provides the means to understand and monitor emotion objectively using automated systems and devices. Many emotion recognition techniques have been explored; however, interpretability of results remains limited but crucial for clinical adoption. This work has presented an analysis of EEG-based emotion recognition across a variety of conditions and use cases.

5.1 Overview

Chapter 1 reviewed the importance of understanding mental health problems and outlined the goals and scope of this work.

In Chapter 2, the emotion classification and affective computing literature was reviewed to provide context about the state of the field. An extensive review of affect databases was conducted, which comprise a variety of physiological modalities and stimulus types and lengths. From this review, the AMIGOS [3] dataset was identified for use in this work.

Chapter 3 extended upon the original AMIGOS work by extracting an additional 112 features and exploring wrapper-style feature selection. Additionally, three lesser-used cross-validation techniques were evaluated to understand the robustness of the models under different combinations of un/known subjects and un/known stimuli.

Consensus was found between the four models, suggesting that HFD features and information from the temporal lobe are important in the classification of emotion from EEG. As the techniques became more generalizable, a performance drop was noted, yet this was marginal for arousal. The exception to this trend was LOMO-Within, which likely suffered as a result of the few movies per participant and the class imbalance.

Chapter 4 evaluated the impact and quality of the labelling strategy used with AMIGOS, along with different methods of training the classifier. Results showed that the dataset included both gender bias and a bias towards low affect labels, and had poor inter-annotator agreement. Gender disaggregated models yielded better performance for women-only models than in the mixed case. Various techniques were explored to reduce the impact of affect bias, and a selective classification scheme was proposed.

5.2 Contribution

The main contributions of this thesis are as follows:

- A review of twenty affective databases was conducted, compiling a comparative resource for affective computing researchers.
- A selection of 217 EEG-based features were compiled from across the literature and used to perform feature selection for emotion classification. Results significantly outperformed the original AMIGOS long movie results, improving F1-scores from 0.557 to 0.706 for valence and from 0.571 to 0.699 for arousal.
- This work corroborated previous findings that the Higuchi Fractal Dimension (HFD) feature contains important discriminative information for EEG emotion recognition, as it was frequently chosen as the top feature across various models.

- A comprehensive analysis of four cross-validation techniques was performed. Each cross-validation scheme evaluated a different use case. Again, results outperformed the baseline AMIGOS results, even when tested with unknown users and stimuli (LOPMO).
- An investigation of the quality of labels in the AMIGOS dataset was conducted, yielding serious limitations in annotator agreement and gender and affect bias.
- The impact of gender bias and differences were explored. Importantly, gender-disaggregated models, particularly female-only models, were found to outperform the mixed models case.
- A selective classification scheme, using rejection of lower confidence decisions, was proposed to improve effective classification performance. This framework could allow future researchers to trade the frequency of decisions for more robust classification decisions.

5.3 Limitations

The exploration of emotion classification is a broad and complex problem. Although a number of investigations were conducted in this work, there are both limitations and potential future work that should be considered. The extensive evaluation of the AMIGOS dataset in this work highlighted the lack of robust publicly available datasets for continuous emotion recognition research. The initial motivation was to explore the temporal dynamics of the evolution of emotion over time. As such, the AMIGOS dataset was chosen because it had both 'long' stimuli and 'continuous' targets using the dimensional affect model. As the work progressed, it was found that this dataset included several limitations, as highlighted throughout this work. Importantly, among them was the use of only three annotators in attempting to infer ground truth emotion from randomly ordered facial expressions. Many of the other

limitations of this dataset were explored, including a limited labelling period (twenty seconds), poor use of the label dynamic range, and affect bias. The former of these severely limited our ability to evaluate the impact of leveraging time-series analysis techniques. The latter, despite attempts to compensate, may limit the generalizability of the results presented here if evaluated on a larger, less biased dataset.

Despite several restrictions incurred by the dataset itself, there were limitations in the work as well. Although 112 additional features were identified and implemented, likely, this is not an exhaustive set. Furthermore, although feature selection was conducted using both SFS and SBS, these are both wrapper-style techniques that rely on a specific classifier model. While SFS is faster to conduct, it is limited by a lack of knowledge about feature correlations. SBS may better incorporate such information but can be prohibitively time-consuming (such as when paired with SVM). Other techniques, such as genetic algorithms, or filter-style methods such as Minimum Redundancy Maximum Relevance (mRMR) could provide better emotion recognition performance and/or added perspective on the types of features chosen. Other classifiers should also be explored as alternatives to the SVM and LDA classifiers used here.

The F1-score was used as the evaluation metric in this work because it provides some level of resilience to class bias in classification problems. Here, it was computed by averaging the F1-score for each partition. While this technique is commonly used in the literature, computing the F1-score based on a global pool of false positives, true positives, and false negatives could change the observed results due to differences in affect label distributions between subjects [173]. Metrics other than F1 could also have been used, such as Youden's J statistic.

The investigation of labels in Chapter 4 was also almost wholly performed using only the LDA classifier with a LOMO-Inter cross-validation scheme. Similarly, for some investigations, the features selected in Chapter 3 were used directly without

recomputing the feature selection process. This was done given the considerable time it took to process these analyses. Although the interpretation of the results remains valid, other classifiers and feature sets may have yielded better performance. Finally, the investigations in Chapter 4 were exploratory, and not necessarily definitive. The investigation of window lengths and increment was severely limited by the temporal resolution of the labels and should be interpreted with caution. Similarly, the performance comparisons related to label bias, annotator differences, and gender bias were all potentially confounded by differences in the amounts of data available for training and/or testing.

5.4 Recommendations for Future Work

An integral aspect of this work is its use of the labels created by the AMIGOS external annotators as ground truth. The impact of these labels was explored but it is infeasible to artificially correct them without introducing additional data-driven bias. Instead, it was recommended that additional annotators be recruited to create more labels directly from the videos provided with the AMIGOS dataset. Other works have also sought to automatically create labels based on facial affect using computer vision techniques [174], which could help to automate this process. The affect labels could also be informed more generally through video content analysis, which has shown to be effective for emotion recognition [175].

Although the AMIGOS dataset has been widely adopted since its publication in 2018, this work has identified that it contains several limitations. Given the potential to learn more about, and to leverage, how emotion evolves, it would be highly beneficial to create a new database designed specifically for this task. Such a database would necessarily include labels with a finer temporal resolution. Creating moment-to-moment annotations with a joystick, for example, has shown promising results [94].

Another potential technique could be to have the individual retrospectively label the data as in [102] or to have a separate individual who knows the participant well label the data. It has been shown that people who are familiar with one another can better quantify the experienced emotion [176].

A new database should also provide more data for each participant, whether that be by including more movies or increasing the movie duration to whole movies. More data would provide more training examples, likely for both classes, reducing bias and the impact of differences in training data for the different evaluation schemes. Potentially, a dataset could initially focus on two opposite genres, such as comedy and suspense, before branching out to others to better understand opposing emotions.

If a new database is not possible, fusion with other datasets could be considered. SEED-VIG [104], for instance, collected EEG from individuals as they drove cars in simulated environments. While there are currently no valence and arousal annotations, facial video is available. If labels were to be created, such as from computer vision, the data from this dataset could potentially be combined with the data from AMIGOS to form a larger set for the LOPMO cross-validation framework.

With improved label frequency, additional features could be explored, such as those focused on temporal information. Other modalities could also be introduced to create a multi-modal system, as multi-modal systems have been shown to improve the emotion recognition performance of models. As discussed above, additional feature selection techniques and classifiers should be explored. Recent works have shown great promise in unsupervised and deep learning techniques [69]. Combining classifiers such as LSTM (which learn temporal information about a signal) with emerging explainable machine learning tools (such as GradCam) could lead to an improved understanding of the temporal nature of emotion using the long films. Other classifiers could also be explored, including XGB or ELM, which have been shown to perform well using the AMIGOS dataset [69, 113].

Bibliography

[1] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[2] A. Martínez-Rodrigo, A. Fernández-Sotos, J. M. Latorre, J. Moncho-Bogani, and A. Fernández-Caballero, "Neural correlates of phrase rhythm: an EEG study of bipartite vs. rondo sonata form," *Frontiers in Neuroinformatics*, vol. 11, p. 29, 2017.

[3] J. A. Miranda-Correa, M. Khomami Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, vol. 3045, no. i, pp. 1–14, 2018.

[4] Mental Health Commission of Canada, "Making the Case for Investing in Mental Health in Canada," tech. rep.

[5] Public Health Agency of Canada, "Mood and anxiety disorders in Canada." https://www.canada.ca/content/dam/canada/health-canada/migration/ healthy-canadians/publications/diseases-conditions-maladies-affections/mental-mood-anxiety-anxieux-humeur/alt/mental-mood-anxiety-anxieux-humeur-eng.pdf, Accessed: 06-12-2019, 2015.

[6] The Government of Canada, *The Human Face of Mental Health and Mental Illness in Canada 2006*. 2006.

[7] A. Sunderland and L. C. Findlay, "Perceived need for mental health care in Canada: Results from the 2012 Canadian community health survey – mental health," *Health Reports*, vol. 24, no. 82, pp. 3–9, 2013.

[8] M. H. C. of Canada, *Mental Health First Aid Canada: For Adults Who Interact With Youth*. Mental Health Commission of Canada, 2010.

[9] E. A. Osuch, E. Vingilis, S. Fisman, and C. Summerhurst, "Early Intervention in Mood and Anxiety Disorders: The First Episode Mood and Anxiety Program (FEMAP).," *Healthcare Quarterly (Toronto, Ont.)*, vol. 18 Suppl, pp. 42–49, 2016.

[10] E. Mower, S. Member, and M. J. Mataric, "A Framework for Automatic Human Emotion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[11] M. F. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database," *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pp. 65–70, 2010.

[12] X. Niu, L. Chen, H. Xie, Q. Chen, and H. Li, "Emotion pattern recognition using physiological signals," *Sensors & Transducers*, vol. 172, no. 6, p. 147, 2014.

[13] C. He, Y.-j. Yao, and X.-s. Ye, "An Emotion Recognition System Based on Physiological Signals Obtained by Wearable Sensors," in *Wearable Sensors and Robots*, pp. 15–25, 2017.

[14] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based emotion recognition and its applications," in *Transactions on Computational Science XII*, pp. 256–277, Springer, 2011.

[15] M. Balconi and C. Lucchiari, "EEG correlates (event-related desynchronization) of emotional face elaboration: a temporal analysis," *Neuroscience Letters*, vol. 392, no. 1-2, pp. 118–123, 2006.

[16] X.-W. Wang, D. Nie, and B.-L. Lu, "EEG-based emotion recognition using frequency domain features and support vector machines," in *International Conference on Neural Information Processing*, pp. 734–743, Springer, 2011.

[17] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors (Switzerland)*, vol. 18, no. 7, p. 2074, 2018.

[18] S. M. Alarcao and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.

[19] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & Emotion*, vol. 9, pp. 87–108, Jan 1995.

[20] A. Schaefer, F. Nils, P. Philippot, and X. Sanchez, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[21] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, I. Elamvazuthi, H. T. Nguyen, and S. Su, "Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-Novem, pp. 7250–7253, 2015.

[22] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[23] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 6, pp. 1052–1063, 2011.

[24] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *8th International Conference on Spoken Language Processing, ICSLP 2004*, pp. 1329–1332, 2004.

[25] E. Campbell, A. Phinyomark, and E. Scheme, "Feature Extraction and Selection for Pain Recognition Using Peripheral Physiological Signals," *Frontiers in Neuroscience*, vol. 13, pp. 1–17, May 2019.

[26] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[27] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[28] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[29] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *International Journal of Psychophysiology*, vol. 61, no. 1, pp. 5–18, 2006.

[30] R. Lane, K. Mcrae, E. Reiman, K. Chen, G. Ahern, and J. Thayer, "Neural correlates of heart rate variability during emotion," *NeuroImage*, vol. 44, pp. 213–222, Jan 2009.

[31] M. A. Yingliang, H. M. Paterson, and F. E. Pollick, "A motion capture library for the study of identity, gender, and emotion perception from biological motion," *Behavior Research Methods*, vol. 38, no. 1, pp. 134–141, 2006.

[32] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 71–82, Springer, 2007.

[33] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals," *IEEE Transactions on Affective Computing*, vol. 9, pp. 550–562, Oct 2018.

[34] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image and Vision Computing*, vol. 25, no. 12, pp. 1856–1863, 2007.

[35] P. J. Lang, M. M. Bradley, B. N. Cuthbert, *et al.*, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.

[36] M. Li and B. L. Lu, "Emotion classification based on gamma-band EEG," *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, pp. 1323–1326, 2009.

[37] S. Basu, N. Jana, A. Bag, M. Mahadevappa, J. Mukherjee, S. Kumar, and R. Guha, "Emotion recognition based on physiological signals using valence-arousal model," *Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015*, no. February 2019, pp. 50–55, 2016.

[38] P. Gong, H. T. Ma, and Y. Wang, "Emotion recognition based on the multiple physiological signals," *2016 IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016*, pp. 140–143, 2016.

[39] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3498–3510, 2012.

[40] H. Bo, L. Ma, Q. Liu, R. Xu, and H. Li, "Music-evoked emotion recognition based on cognitive principles inspired EEG temporal and spectral features," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 9, pp. 2439–2448, 2019.

[41] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG Pattern Analysis for Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 3, pp. 102–115, Jan 2012.

[42] P. J. Bota, C. Wang, A. L. Fred, and H. Placido Da Silva, "A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019.

[43] M. Ali, F. Al Machot, A. H. Mosa, M. Jdeed, E. Al Machot, and K. Kyamakya, "A globally generalized emotion recognition system involving different physiological signals," *Sensors (Switzerland)*, vol. 18, no. 6, pp. 1–19, 2018.

[44] P. Ekman, "An Argument for Basic Emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[45] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[46] D. Plass-Oude Bos, "EEG-based Emotion Recognition: The Influence of Visual and Auditory Stimuli," 2006.

[47] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian, "Recognition of Advertisement Emotions with Application to Computational Advertising," *IEEE Transactions on Affective Computing*, 2020.

[48] S. Koelstra, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Y. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[49] C. L. Lisetti and F. Nasoz, "Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals," *Eurasip Journal on Applied Signal Processing*, vol. 2004, no. 11, pp. 1672–1687, 2004.

[50] D. Palomba, M. Sarlo, a. Angrilli, a. Mini, and L. Stegagno, "Cardiac Responses Associated with Affective Processing of Unpleasant Film Stimuli," *International Journal of Psychophysiology*, vol. 36, pp. 45–57, 2000.

[51] J. Healey, "Recording Affect in the Field: Towards Methods and Metrics for Improving Ground Truth Labels," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6974, no. PART 1, pp. 107–116, 2011.

[52] A. Hanjalic and L. Q. Xu, "Affective Video Content Representation and Modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[53] B. Kolb and I. Whishaw, *Fundamentals of Human Neuropsychology*. NY: Worth Publishers, 7 ed., 2015.

[54] V. Doma and M. Pirouz, "A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals," *Journal of Big Data*, vol. 7, no. 1, pp. 1–21, 2020.

[55] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[56] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *International Workshop on Multimedia Content Representation, Classification and Security*, pp. 530–537, Springer, 2006.

[57] K. Takahashi and A. Tsukaguchi, "Remarks on emotion recognition from multimodal bio-potential signals," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, vol. 2, pp. 1654–1659, IEEE, 2003.

[58] Z. Khalili and M. Moradi, "Emotion detection using brain and peripheral signals," in *2008 Cairo International Biomedical Engineering Conference*, pp. 1–4, IEEE, 2008.

[59] M. Soleymani, M. Pantic, and T. Pun, "Multimodal Emotion Recognition in Response to Videos," *IEEE Transactions on Affective Computing*, vol. 3, pp. 211–223, Apr 2012.

[60] W. L. Zheng, B. N. Dong, and B. L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*, pp. 5040–5043, 2014.

[61] O. Barral, I. Kosunen, and G. Jacucci, "No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment," *ACM Transactions on Computer-Human Interaction*, vol. 24, no. 6, pp. 1–29, 2017.

[62] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized emotion recognition by personality-aware high-order learning of physiological signals," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 15, no. 1s, pp. 1–18, 2019.

[63] Y. Zhao, X. Cao, J. Lin, D. Yu, and X. Cao, "Multimodal Emotion Recognition Model using Physiological Signals," *arXiv preprint arXiv:1911.12918*, 2019.

[64] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors," *IEEE Transactions on Affective Computing*, vol. 9, pp. 147–160, Apr 2018.

[65] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm," in *2007 IEEE International Conference on Systems, Man and Cybernetics*, pp. 2662–2667, IEEE, 2007.

[66] Siddharth, T.-P. Jung, and T. J. Sejnowski, "Multi-modal Approach for Affective Computing," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 291–294, IEEE, July 2018.

[67] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[68] A. Yazdani, J. S. Lee, J. M. Vesin, and T. Ebrahimi, "Affect recognition based on physiological changes during the watching of music videos," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 1, pp. 1–26, 2012.

[69] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing," *IEEE Transactions on Affective Computing*, pp. 1–12, 2019.

[70] R. Adolphs, D. Tranel, and A. R. Damasio, "Dissociable neural systems for recognizing emotions," *Brain and Cognition*, vol. 52, no. 1, pp. 61–69, 2003.

[71] A. Bhardwaj, A. Gupta, P. Jain, A. Rani, and J. Yadav, "Classification of human emotions from EEG signals using SVM and LDA Classifiers," in *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 180–185, IEEE, 2015.

[72] M. Stikic, R. R. Johnson, V. Tan, and C. Berka, "EEG-based classification of positive and negative affective states," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 99–112, 2014.

[73] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[74] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, 2018.

[75] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[76] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 186–197, 2009.

[77] H. Xu and K. N. Plataniotis, "Affect recognition using EEG signal," in *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 299–304, IEEE, 2012.

[78] P. Ackermann, C. Kohlschein, J. Á. Bitsch, K. Wehrle, and S. Jeschke, "EEG-based automatic emotion recognition: Feature extraction, selection and classification methods," *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services, Healthcom 2016*, pp. 1–6, 2016.

[79] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring EEG features in cross-subject emotion recognition," *Frontiers in Neuroscience*, vol. 12, p. 162, 2018.

[80] A. Gupta, H. Sahu, N. Nanecha, P. Kumar, P. P. Roy, and V. Chang, "Enhancing text using emotion detected from EEG signals," *Journal of Grid Computing*, vol. 17, no. 2, pp. 325–340, 2019.

[81] S. Hatamikia and A. M. Nasrabadi, "Recognition of emotional states induced by music videos based on nonlinear feature extraction and som classification," in *2014 21th Iranian Conference on Biomedical Engineering (ICBME)*, pp. 333–337, IEEE, 2014.

[82] D.-W. Chen, R. Miao, W.-Q. Yang, Y. Liang, H.-H. Chen, L. Huang, C.-J. Deng, and N. Han, "A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition," *Sensors*, vol. 19, no. 7, p. 1631, 2019.

[83] S. Solhjoo, A. M. Nasrabadi, and M. R. H. Golpayegani, "EEG-based mental task classification in hypnotized and normal subjects," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 2041–2043, IEEE, 2006.

[84] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.

[85] E.-J. Chang, A. Rahimi, L. Benini, and A.-Y. A. Wu, "Hyperdimensional Computing-based Multimodality Emotion Recognition with Physiological Signals," *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 137–141, 2019.

[86] T. Althobaiti, S. Katsigiannis, D. West, and N. Ramzan, "Examining Human-Horse Interaction by Means of Affect Recognition via Physiological Signals," *IEEE Access*, vol. 7, no. June, pp. 77857–77867, 2019.

[87] A. Martínez-Rodrigo, B. García-Martínez, L. Zunino, R. Alcaraz, and A. Fernández-Caballero, "Multi-lag analysis of symbolic entropies on EEG recordings for distress recognition," *Frontiers in Neuroinformatics*, vol. 13, no. June, pp. 1–15, 2019.

[88] P. Lakhan, N. Banluesombatkul, V. Changniam, R. Dhithijaiyratn, P. Leelaarporn, E. Boonchieng, S. Hompoonsup, and T. Wilaiprasitporn, "Consumer Grade Brain Sensing for Emotion Recognition," *IEEE Sensors Journal*, pp. 1–1, 2019.

[89] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H. C. Traue, P. Werner, A. Al-Hamadi, N. Diniz, G. M. da Silva, and A. O. Andrade, "Automatic pain quan-

tification using autonomic parameters," *Psychology and Neuroscience*, vol. 7, no. 3, pp. 363–380, 2014.

[90] Y. Dai, X. Wang, X. Li, and P. Zhang, "Reputation-driven multimodal emotion recognition in wearable biosensor network," *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, vol. 2015-July, pp. 1747–1752, 2015.

[91] Y.-H. Liu, C.-T. Wu, W.-T. Cheng, Y.-T. Hsiao, P.-M. Chen, and J.-T. Teng, "Emotion recognition from single-trial EEG based on kernel Fisher's emotion pattern and imbalanced quasiconformal kernel support vector machine," *Sensors*, vol. 14, no. 8, pp. 13361–13388, 2014.

[92] R. M. Mehmood, R. Du, and H. J. Lee, "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors," *IEEE Access*, vol. 5, pp. 14797–14806, 2017.

[93] J. A. Miranda-Correa and I. Patras, "A multi-task cascaded network for prediction of affect, personality, mood and social context using EEG signals," *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pp. 373–380, 2018.

[94] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 7, pp. 17–28, Jan 2016.

[95] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H. C. Traue, and S. Gruss, "'BioVid Emo DB': A Multimodal Database for Emotion Analyses validated by Subjective Ratings," *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–6, 2016.

[96] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[97] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 98–107, Jan 2018.

[98] S. Carvalho, J. Leite, S. Galdo-Álvarez, and Ó. F. Gonçalves, "The emotional movie database (EMDB): A self-report and psychophysiological study," *Applied Psychophysiology Biofeedback*, vol. 37, no. 4, pp. 279–294, 2012.

[99] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159–170, 2017.

[100] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[101] A. Savran, K. Ciftci, G. Chanel, J. C. Mota, L. H. Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut, "Emotion Detection in the Loop from Brain Signals and Facial Images," in *Proceedings of the eNTERFACE 2006 Workshop*, 2006.

[102] S. Schneegass, B. Pfleging, N. Broy, A. Schmidt, and F. Heinrich, "A data set of real world driving to assess driver workload," *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2013*, pp. 150–157, 2013.

[103] W.-L. Zheng and B.-L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, pp. 162–175, Sept 2015.

[104] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A Multimodal Framework for Recognizing Human Emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.

[105] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, 2017.

[106] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The Swell knowledge work dataset for stress and user modeling research," *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, pp. 291–298, 2014.

[107] E. Vyzas and R. W. Picard, "Offline and online recognition of emotion expression from physiological data," *Workshop on Emotion-Based Agent Architectures at the Third International Conference on Autonomous Agents*, no. 488, pp. 135–142, 1999.

[108] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Introducing WeSAD, a multimodal dataset for wearable stress and affect detection," *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pp. 400–408, 2018.

[109] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'Trier social stress test' - A tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[110] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2376–2379, IEEE, May 2011.

[111] Y. Baveye, E. Dellandrea, C. Chamaret, and Liming Chen, "LIRIS-ACCEDE: A Video Database for Affective Content Analysis," *IEEE Transactions on Affective Computing*, vol. 6, pp. 43–55, Jan 2015.

[112] M. Soleymani, J. J. Kierkels, G. Chanel, and T. Pun, "A Bayesian framework for video affective representation," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, (Amsterdam, Netherlands), pp. 1–7, IEEE, Sept 2009.

[113] K. Tung, P. K. Liu, Y. C. Chuang, S. H. Wang, and A. Y. Wu, "Entropy-assisted multi-modal emotion recognition framework based on physiological signals," *2018 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2018 - Proceedings*, pp. 22–26, 2019.

[114] M. Gjoreski, M. Luštrek, M. Gams, and B. Mitrevski, "An inter-domain study for arousal recognition from physiological signals," *Informatica (Slovenia)*, vol. 42, no. 1, pp. 61–68, 2018.

[115] D. Hutchison and J. C. Mitchell, "Transactions on Computational Science XVII," vol. 1, pp. 172–185, 1973.

[116] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity," *IEEE Transactions on Affective Computing*, vol. 3045, 2019.

[117] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Transactions on Multimedia*

Computing, Communications, and Applications (TOMM), vol. 15, no. 2, pp. 1–23, 2019.

[118] S. Ganesh, A. M. Chinchani, A. Bhushan, D. Kanchan, and S. Kubakaddi, "Participant-dependent and participant-independent classification of emotions using EEG signals," *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 357–364, 2017.

[119] L. Bozhkov, P. Georgieva, I. Santos, A. Pereira, and C. Silva, "EEG-based subject independent affective computing models," *Procedia Computer Science*, vol. 53, no. 1, pp. 375–382, 2015.

[120] S. Kinreich, A. Djalovski, L. Kraus, Y. Louzoun, and R. Feldman, "Brain-to-Brain Synchrony during Naturalistic Social Interactions," *Scientific Reports*, vol. 7, Dec 2017.

[121] E. J. Boothby, M. S. Clark, and J. A. Bargh, "Shared Experiences Are Amplified," *Psychological Science*, vol. 25, no. 12, pp. 2209–2216, 2014.

[122] K. Schaaff and T. Schultz, "Towards emotion recognition from electroencephalographic signals," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6, IEEE, 2009.

[123] L. Tian, M. Muszynski, C. Lai, J. D. Moore, T. Kostoulas, P. Lombardo, T. Pun, and G. Chanel, "Recognizing induced emotions of movie audiences: Are induced and perceived emotions the same?," *2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 28–35, 2017.

[124] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm, "An affective computing approach to physiological emotion specificity:

Toward subject-independent and stimulus-independent classification of film-induced emotions," *Psychophysiology*, vol. 48, no. 7, pp. 908–922, 2011.

[125] S. H. Wang, H. T. Li, E. J. Chang, and A. Y. Andy Wu, "Entropy-assisted emotion recognition of valence and arousal using XGBoost classifier," *IFIP Advances in Information and Communication Technology*, vol. 519, no. May, pp. 249–260, 2018.

[126] H. C. Yang and C. C. Lee, "Annotation Matters: A Comprehensive Study on Recognizing Intended, Self-reported, and Observed Emotion Labels using Physiology," *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, pp. 413–419, 2019.

[127] H.-C. Yang and C.-C. Lee, "An attribute-invariant variational learning for emotion recognition using physiology," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1184–1188, IEEE, 2019.

[128] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Information Processing and Management*, vol. 57, no. 3, p. 102185, 2020.

[129] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar, "Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019.

[130] R. Harper and J. Southern, "A Bayesian Deep Learning Framework for End-To-End Prediction of Emotion from Heartbeat," *IEEE Transactions on Affective Computing*, 2020.

[131] R. Harper and J. Southern, "End-To-End Prediction of Emotion From Heartbeat Data Collected by a Consumer Fitness Tracker," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–7, IEEE, 2019.

[132] P. Sarkar and A. Etemad, "Self-supervised learning for ECG-based emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3217–3221, IEEE, 2020.

[133] https://www.emotiv.com/epoc/.

[134] G. Gómez-Herrero, K. Rutanen, and K. Egiazarian, "Blind source separation by entropy rate minimization," *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 153–156, 2009.

[135] R. W. Homan, J. Herman, and P. Purdy, "Cerebral location of international 10–20 system electrode placement," *Electroencephalography and Clinical Neurophysiology*, vol. 66, no. 4, pp. 376–382, 1987.

[136] F. S. Bao, X. Liu, and C. Zhang, "PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction," *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–7, 2011.

[137] R. Vallat, "EntroPy." https://github.com/raphaelvallat/entropy, Accessed: 02-20-2019, 2018.

[138] O. Dressler, G. Schneider, G. Stockmanns, and E. Kochs, "Awareness and the EEG power spectrum: analysis of frequencies," *British Journal of Anaesthesia*, vol. 93, no. 6, pp. 806–809, 2004.

[139] M. Zangeneh Soroush, K. Maghooli, S. K. Setarehdan, and A. Motie Nasrabadi, "A Review on EEG Signals Based Emotion Recognition," *International Clinical Neuroscience Journal*, vol. 4, no. 4, pp. 118–129, 2017.

[140] L. Aftanas, A. Varlamov, S. Pavlov, V. Makhnev, and N. Reva, "Event-related synchronization and desynchronization during affective processing: Emergence of valence-related time-dependent hemispheric asymmetries in theta and upper alpha band," *International Journal of Neuroscience*, vol. 110, no. 3-4, pp. 197–219, 2001.

[141] W. Klimesch, M. Doppelmayr, H. Russegger, T. Pachinger, and J. Schwaiger, "Induced alpha band power changes in the human EEG and attention," *Neuroscience Letters*, vol. 244, no. 2, pp. 73–76, 1998.

[142] D. J. Oathes, W. J. Ray, A. S. Yamasaki, T. D. Borkovec, L. G. Castonguay, M. G. Newman, and J. Nitschke, "Worry, Generalized Anxiety Disorder, and Emotion: Evidence from the EEG gamma band," *Biological Psychology*, vol. 79, no. 2, pp. 165–170, 2008.

[143] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cognition & Emotion*, vol. 15, no. 4, pp. 487–500, 2001.

[144] D. J. Schutter, P. Putman, E. Hermans, and J. van Honk, "Parietal electroencephalogram beta asymmetry and selective attention to angry facial expressions in healthy human subjects," *Neuroscience Letters*, vol. 314, no. 1-2, pp. 13–16, 2001.

[145] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306–310, 1970.

[146] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 82–87, 1995.

[147] S. Sanyal, A. Banerjee, R. Pratihar, A. K. Maity, S. Dey, V. Agrawal, R. Sengupta, and D. Ghosh, "Detrended Fluctuation and Power Spectral Analysis of alpha and delta EEG brain rhythms to study music elicited emotion," in *2015 International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 205–210, IEEE, 2015.

[148] C. Goh, B. Hamadicharef, G. Henderson, and E. Ifeachor, "Comparison of fractal dimension algorithms for the computation of EEG biomarkers for dementia," in *2nd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2005)*, 2005.

[149] S. Kesić and S. Z. Spasić, "Application of Higuchi's fractal dimension from basic to clinical neurophysiology: A review," *Computer Methods and Programs in Biomedicine*, vol. 133, pp. 55–70, 2016.

[150] S. J. Roberts, W. Penny, and I. Rezek, "Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing," *Medical & Biological Engineering & Computing*, vol. 37, no. 1, pp. 93–98, 1999.

[151] C. J. James and D. Lowe, "Extracting multisource brain activity from a single electromagnetic channel," *Artificial Intelligence in Medicine*, vol. 28, no. 1, pp. 89–104, 2003.

[152] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn:

Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[153] D. Girardi, F. Lanubile, and N. Novielli, "Emotion detection using noninvasive low cost sensors," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 125–130, IEEE, 2017.

[154] F. Song, D. Mei, and H. Li, "Feature Selection Based on Linear Discriminant Analysis," in *2010 International Conference on Intelligent System Design and Engineering Application*, vol. 1, pp. 746–749, IEEE, 2010.

[155] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[156] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[157] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[158] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[159] V. H. Anh, M. N. Van, B. B. Ha, and T. H. Quyet, "A real-time model based support vector machine for emotion recognition through EEG," in *2012 Interna-*

tional Conference on Control, Automation and Information Sciences (ICCAIS), pp. 191–196, IEEE, 2012.

[160] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, "Real-time EEG-based happiness detection system," *The Scientific World Journal*, vol. 2013, 2013.

[161] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Tutorial and research workshop on affective dialogue systems*, pp. 36–48, Springer, 2004.

[162] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, *et al.*, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[163] A. Ali, S. M. Shamsuddin, A. L. Ralescu, *et al.*, "Classification with class imbalance problem: A review," *International Journal of Advances in Soft Computing and its Applications*, vol. 7, no. 3, pp. 176–204, 2015.

[164] G. G. Knyazev, J. Y. Slobodskoj-Plusnin, and A. V. Bocharov, "Gender differences in implicit and explicit processing of emotional facial expressions as revealed by event-related theta synchronization.," *Emotion*, vol. 10, no. 5, p. 678, 2010.

[165] J.-Y. Zhu, W.-L. Zheng, and B.-L. Lu, "Cross-subject and cross-gender emotion classification from EEG," in *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*, pp. 1188–1191, Springer, 2015.

[166] T. Lee, H. Liu, C. Chan, S. Fang, and J. Gao, "Neural activities associated with emotion recognition observed in men and women," *Molecular Psychiatry*, vol. 10, no. 5, pp. 450–455, 2005.

[167] C. Crivelli, S. Jarillo, J. A. Russell, and J.-M. Fernández-Dols, "Reading emotions from faces in two indigenous societies.," *Journal of Experimental Psychology: General*, vol. 145, no. 7, p. 830, 2016.

[168] A. Schützwohl and R. Reisenzein, "Facial expressions in response to a highly surprising event exceeding the field of vision: A test of Darwin's theory of surprise," *Evolution and Human Behavior*, vol. 33, no. 6, pp. 657–664, 2012.

[169] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[170] Y.-H. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, 2010.

[171] S. Ovadia, "Ratings and rankings: Reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, 2004.

[172] P. A. Russell and C. D. Gray, "Ranking or rating? some data and their implications for the measurement of evaluative response," *British Journal of Psychology*, vol. 85, no. 1, pp. 79–92, 1994.

[173] G. Forman and M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement," *ACM SIGKIDD Explorations Newsletter*, vol. 12, no. 1, pp. 49–57, 2010.

[174] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition.," *IEEE Transac-*

tions on Pattern Analysis & Machine Intelligence, vol. 37, no. 6, pp. 1113–1133, 2015.

[175] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, pp. 1–26, Mar 2018.

[176] J. T. Stanley and D. M. Isaacowitz, "Caring More and Knowing More Reduces Age-Related Differences in Emotion Perception," *Psychology and Aging*, vol. 30, no. 2, p. 383, 2015.

Vita

Candidate's full name: Nicole Bendrich
University attended: BScEE, University of New Brunswick, 2015
Publications: N/A
Conference Presentations: N/A